IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 26, 2019, accepted May 9, 2019, date of publication June 10, 2019, date of current version August 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922148

Evolving Coauthorship Modeling and Prediction via Time-Aware Paired Choice Analysis

LIANG HU[1,2], QINGKUI CHEN[1], LONGBING CAO[2], SONGLEI JIAN[3], HAIYAN ZHAO[1], AND JIAN CAO[4]

[1]Department of Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2]Advanced Analytics Institute, University of Technology Sydney, Sydney, NSW 2008, Australia
[3]College of Computer, National University of Defense Technology, Changsha 410073, China
[4]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Qingkui Chen (chenqingkui@usst.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572325 and Grant 61472253, in part by the Shanghai Key Programs of Science and Technology under Grant 16DZ1203603, in part by the Shanghai Engineering Research Center Project under Grant GCZX14014 and Grant C14001, and in part by the National Key Research and Development Plan under Grant 2018YFB1003800.

ABSTRACT Coauthorship prediction is challenging yet important for academic collaboration and novel research topics discovery. The challenges lie in the dynamics of social or organizational relationships, changing preferences of suitable collaborators, and the evolution of research interests or topics. However, most current approaches and systems developed so far are mainly based on past coauthorships from a static viewpoint and do not capture the above evolving characteristics in coauthoring. Accordingly, this paper proposes a time-aware approach to capture the evolving coauthorships from online academic databases in terms of capturing the dynamics of social relationships and research interests. In particular, in order to understand the underlying factors influencing researchers to make choices of coauthors, we incorporate choice modeling based on utility theory. More specifically, our model conducts a series of pairwise choices over a poset induced by a utility function so as to learn the preference over all candidate coauthors. To complete the model inference, a gradient-based algorithm is devised to efficiently learn the model parameters for large-scale data. Finally, extensive experiments conducted on a real-world dataset show that our approach consistently outperforms other state-of-the-art methods.

INDEX TERMS Discrete choice modeling, dynamic social network, temporal link analysis, utility theory.

I. INTRODUCTION

Nowadays, quite a few online academic databases have been built to access bibliographic information. In particular, such databases are mainly used for finding and accessing articles in academic archives. However, coauthorship is another important information that can be further exploited from these academic databases. For example, a researcher often desires to find suitable collaborators for a specific task. Although we can manually browse and select researchers with the aid of databases, an automated retrieval system is more desirable to speed up this time-consuming process. Moreover, if we can predict future coauthorships among leading researchers, we may capture newly emerging research topics in advance. Therefore, accurately predicting the future coauthorships in terms of analyzing the information from online academic databases is of practical significance.

Coauthorship prediction aims to find the most possible coauthors whereby, given a researcher alongside his/her profile, e.g., past publication records. Intuitively, historical coauthoring records can provide an informative reference to predict future coauthorships. Hence, some early attempts adopt community detection approach to find close researchers [1], where it assumes that members in the same community have stronger links than those in other communities, so researchers in the same community tend to be first-choice coauthors. Although this approach is helpful, it only considered the topology of a network. Moreover, the successes of coauthor choices are heavily dependent on having compatible collaboration topics as well. Two researchers are unlikely to be coauthors if they are not interested in each other's research fields. *CollabSeer* [2] offers

The associate editor coordinating the review of this manuscript and approving it for publication was Yangming Li.

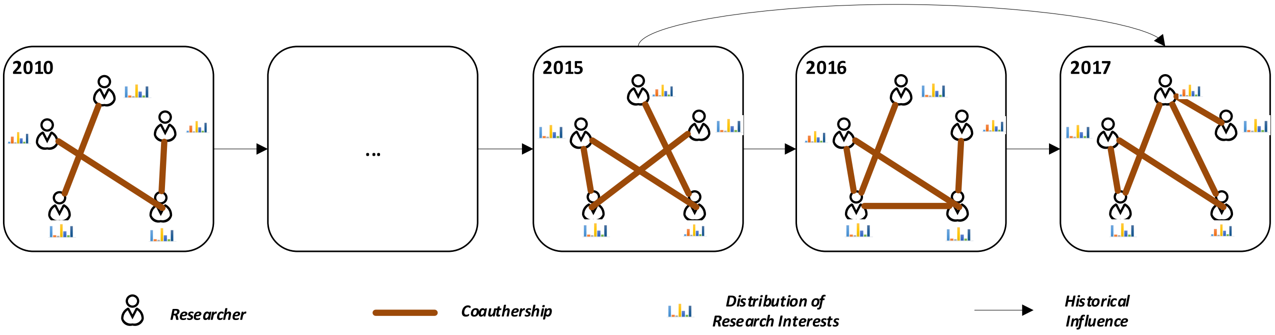


FIGURE 1. The evolution of co-authorships over time. Such evolution is dependent on both historical social relationships and research interests. In fact, social relationships and collaboration topics coevolve over time.

some improvements, predicting coauthorships by jointly modeling the structural similarity over the coauthorship network and the content similarity over the textual material. Tang *et al.* [3] studied a cross-domain collaboration recommendation approach to meet the needs of interdisciplinary cooperation, where the authors explore some methods to find compatible collaboration topics across two related domains using a topic model. Then, they use random walks with restarts (RWR) to measure the relatedness of two researchers.

However, in the real world, coauthorships are constantly evolving. For example, it does not guarantee that two researchers, who have previously coauthored papers, will continue coauthoring more papers in the following years due to some particular reasons, e.g., employment change. Moreover, the formation of coauthorships is heavily dependent on the research interests of the authors. Since the research interests of each author shift over time, in turn, it leads to developing new coauthorships on active research focus. In a word, collaborative relationships and research interests are not static but coevolve over time. Figure 1 demonstrates such phenomena by a series of yearly time slices, where each slice depicts the snapshot of coauthorships over a group of researchers and the distribution of research interests w.r.t. each author. The real-world data also illustrates these observations, cf. Figure 4, the dataset collected for experiments, where researcher A's most significant coauthor changed from #3 (2000-2001), to #1 (2002-2003), then to #2 (2007-2008), while his/her research interest on topic #46 decreased gradually to zero by 2009. In conclusion, there are two primary evolving factors driving the formation of coauthorships: (1) *Social Relationships* and (2) *Research Interests*. However, most current approaches, including the aforementioned ones, studied coauthorships from a static perspective, which are incapable of capturing the dynamic factors of coauthoring. As a result, these methods may fail to accurately capture the evolving coauthorships since they fully make the prediction based on a static state.

To understand and capture the evolution of coauthorships, we first need to model the underlying factors that result in the decision on choosing which coauthors. Fortunately, choice modeling [4] can be employed to analyze the user decision based on the given attributes. In fact, choice modeling is a very powerful tool which has been effectively applied in many areas, including psychology, economics, policy, labor, health, marketing, over decades [4]. In particular, choice modeling is often interpreted through utility theory [5], i.e. users tend to choose the items with high utility. Accordingly, we can measure the utility of each choice on coauthors, so coauthorship prediction is reduced to an optimization problem of choosing coauthors with the maximum utility from the perspective of choice modeling. As a result, predicting the future coauthors w.r.t. a target researcher is equivalent to computing which choices of coauthors can produce maximum utility at that time.

Inspired by the above discussion, we design a time-aware approach to capture the evolving coauthorships. First, we define an order relation of preference over a set of candidate coauthors w.r.t. each year in terms of a utility function. Based on such a preference order, we propose the time-aware paired choice analysis (TAPCA) on coauthorships to model temporal preferences by a series of pairwise comparisons. More specifically, our approach takes the time-varying features of social relationships and collaborative topics as inputs to model the choices of coauthors over time. In addition, to improve the generalization ability of our model, we respectively regularize the parameters of our models with the L1 and L2 norms. Since the total number of pairwise comparisons may be large, standard gradient methods cannot work well. We propose to learn the parameters by the stochastic gradient method, that is, we update the parameters w.r.t. a single choice a time. Once the parameters are estimated, the most possible future coauthorships are ranked by the predicted utility of choices over candidate coauthors.

The main contributions of this paper can be summarized as follows:

- We propose a time-aware approach to capture the dynamic nature of coauthoring in a realistic way.
- We define two aspects of time-varying features: social factors and collaboration topics, to better understand the choices over coauthors with these features.
- We design a choice model over ordered preference relationships, which statistically relates the choices of

coauthors to the time-varying features to learn the influence of each feature.

- We conduct extensive experiments on a real-world academic dataset which shows the superiority of our approach over the other comparison methods.

II. RELATED WORK

To study collaborations, a classic method is to analyze the structure of a collaboration network [6]. Newman [1] proposed a community detection algorithm over a coauthorship network, where a community accommodates people with common research interests. The underlying assumption of predicting coauthorships is that people tend to collaborate with those in the same community. CollabSeer [2] defines the various vertex and lexical similarities to measure the tightness of two researchers, so it can rank potential coauthors by integrating vertex and lexical similarities. Coauthorship prediction can also be viewed from the classic link prediction perspective. Miller *et al.* [7] proposed a probabilistic model to predict links between objects according to the interaction between their latent features. Backstrom and Leskovec [8] integrated features of nodes and edges into a unified model to predict the weight of unseen links using random walks. Makarov *et al.* [9] use the network to power a coauthor recommender system. The system gives recommendations of authors that have interests similar to the target author or whose coauthorship pattern is similar to that of the author. However, these classical methods take a static perspective of a network so they cannot work well on the evolving coauthorships as studied in this paper. Most importantly, these methods neither consider the dynamics of social relationships nor the compatibility of collaboration topics that essentially decide the choice of coauthors.

Sun *et al.* [10], [11] study the problem of future coauthorship prediction in a heterogeneous bibliographic network which contains multiple types of objects, such as authors, venues, topics, and papers, as well as multiple types of links denoting different relations among these objects. Cross-domain Topic Learning (CTL) [3] aims to study the collaboration pattern across domains, where each author needs to be assigned to a domain and models are constructed for each pair of domains in advance. As stated by its name, CTL is a topic model which mainly focuses on finding the patterns of collaboration topics across two domains instead of the patterns of the formation of coauthorships, so it studies a quite different problem from this paper. Moreover, CTL does not explicitly model the influence of social relationships on choosing coauthors; instead, it ranks coauthors by random walks with restart (RWR) where the transition probability is induced from a common topic distribution related to two researchers. Furthermore, the above models do not consider the dynamic nature of collaboration. Now, let us consider an example: a researcher has recently changed his interests but has a large number of old publications, hence the proportion of old research interests may overwhelm the new ones. In such a case, the ranking of potential coauthors is mainly determined according to the old interests so it will lead to a poor prediction.

Nowadays, researchers have realized that coauthorship is an evolving relation [12] that cannot be well captured in terms of static modeling. ACRec [13] is an Academic RWR model using three academic metrics as basics for recommending collaboration. Therein, the latest collaboration time point is one of the metrics are exploited to define the link importance. Acar *et al.* [14] considered the temporal link prediction problem and modeled the evolving link data in terms of tensor factorization. This method only represents links, i.e. whether a link exists between two nodes, but it cannot model the underlying factors resulting in the formation of links, e.g. research interests for each author. TensorCase [15] is a coupled tensor factorization method for forecasting time-evolving networks, which is able to incorporate multiple information sources. The model proposed in this paper is also constructed in a time-aware manner, which can better capture the time-varying factors that determine the formation of collaborations, i.e. a drift in social relationships and collaboration topics.

III. PRELIMINARIES

Since coauthorships are changing over time, we need to represent the dynamic factors influencing coauthoring so as to capture the underlying trends to predict future coauthoring in a more adaptive way. In fact, people are connected by similar research interests in the real world, so closer social relationships may partially imply more suitable research topics for collaboration.

A. TIME-VARYING FEATURES IN COAUTHORING

In the real world, social relationships and research focus are the two most primary aspects that lead to the formation of coauthorships. Since both social relationships and research interests change with time, we need to define these dynamic features.

1) TIGHTNESS OF SOCIAL RELATIONSHIPS

Similar to the way that people trust their friends more than those who are unfamiliar, a researcher also tends to collaborate with those coauthors who have a tight relationship, because collaboration with unacquainted persons means more uncertainties, which may increase the incompatibility of collaboration and the risk of failure.

We denote all researchers for study as $\mathcal{U} = \{u_1, u_2, \cdots, u_N\}$. Now, let's consider a set of coauthorship networks $\left\{G^t\right\}$ organized by time slice t, where each node stands for a researcher $u \in \mathcal{U}$, the weight of edge w_{ij}^t indicates the number of coauthoring between u_i and u_j during t and $\Gamma^t(u_i)$ denotes the neighbors of u_i. Then, we can define the following time-varying social features that may lead to the collaboration between two researchers, denoted as:

$$\mathbf{s}_{ij}^t = \{\#col_{ij}^t, \#col_{ij}^{[t-b,t]}, \#coc_{ij}^t, \#coc_{ij}^{[t-b,t]}, \\ 1(d_{ij}^t \le k), 1(d_{ij}^{[t-b,t]} \le k)\} \qquad (1)$$

TABLE 1. Dynamic social features in collaboration.

$\#col_{ij}^{t}$	The number of observed coauthoring between u_i and u_j during t.
$\#col_{ij}^{[t-b,t]}$	The cumulative number of observed coauthoring between u_i and u_j during $t-b$ to t, where b is a parameter.
$\#coc_{ij}^{t}$	The number of observed common coauthors between u_i and u_j during t
$\#coc_{ij}^{[t-b,t]}$	The cumulative number of observed common coauthors between u_i and u_j during $t-b$ to t, where b is a parameter.
$1(d_{ij}^{t} \leq k)$	Boolean value to indicate if the distance d_{ij} between u_i and u_j is $d_{ij} \leq k$ during t.
$1(d_{ij}^{[t-b,t]} \leq k)$	Boolean value to indicate if the distance d_{ij} between u_i and u_j is $d_{ij} \leq k$ during $t-b$ to t, where b is a parameter.

The description of each feature is given by Table 1. Obviously, all the defined features in Table 1 reflect the tightness between two coauthors in the social network. Therein, we model both the short term factors, e.g. $\#col_{ij}^{t}$ and $\#coc_{ij}^{t}$, and the long term factors, e.g. $\#col_{ij}^{[t-b,t]}$ and $\#coc_{ij}^{[t-b,t]}$. Typically, we set the time period t to be one natural year in this paper. Forexample, we set $t = 2016$ and $b = 2$, and then we obtain $\#col_{ij}^{[2016,2018]}$ which means the number of collaborations between u_i and u_j during 2016 to 2018.

2) COMPATIBILITY OF COLLABORATION TOPICS

Another important factor that determines the formation of collaboration is the research interests of coauthors. It is unlikely that two researchers will engage in collaboration if they are not interested in the research topic of each other. Therefore, we should extract the preferred collaboration topics of each researcher.

Here, we employ the Latent Dirichlet Allocation (LDA) [16], which is a widely used topic model, to extract the distribution of research topics in which each researcher is involved. In LDA, the topic proportions vectors of a document d is denoted as $\boldsymbol{\theta}_d$ where the proportion of each topic z is given by:

$$\theta_{d,z} = \frac{C_{d,z} + \alpha}{\sum_{z'} C_{d,z'} + Z\alpha} \tag{2}$$

where $C_{d,z}$ denotes the count of words assigned to the topic z in d, Z is the total number of topics and α is a hyperparameter. If we denote $N_d = \sum_{z'} C_{d,z'}$ as the length of the document d and set $\alpha = 0$, then we can rewrite Eq. 2 as empirical topic proportions θ_d [17]:

$$\theta_{d,z} = \frac{C_{d,z}}{N_d} \tag{3}$$

With some minor modification, we can easily obtain the statistics of the collaboration topic distribution of a researcher u_i during $t-b$ to t. Let $\mathcal{D}_i^{[t-b,t]}$ denote the documents published during $t-b$ to t, where u_i is a coauthor of each document $d \in \mathcal{D}_i^t$, formally

$$\mathcal{D}_i^{[t-b,t]} = \{d | u_i \in coauthors\,(d) \wedge pubdate\,(d) \in [t-b, t]\} \tag{4}$$

Then, the empirical topic proportions $\boldsymbol{\theta}_i^t$ based on u_i's involved collaboration during $t-b$ to t can be easily obtained:

$$\theta_{i,z}^t = \frac{\sum_{d \in \mathcal{D}_i^{[t-b,t]}} C_{d,z}}{\sum_{d \in D_i^{[t-b,t]}} N_d} \tag{5}$$

Here, the assumption is that the coauthored documents can reflect the collaboration focus of coauthors. For example, u_i has expertise in Machine Learning (ML) and he is also interested in Social Network (SN) but with less expertise, so he more probably tends to coauthor papers of the topics on both ML and SN. As a result, both $\theta_{i,z=ML}^t$ and $\theta_{i,z=SN}^t$ tend to have higher proportions in his coauthored papers, which reflect the focus of u_i's collaboration.

Since the topic distribution θ_i^t can reflect u_i's collaboration focus, u_i is more likely to collaborate with whomever has some common interest. Therefore, it is possible to measure the compatibility of collaboration topics between two researchers u_i and u_j in terms of $\boldsymbol{\theta}_i^t$ and $\boldsymbol{\theta}_j^t$. For example, we can compute the normalized Hadamard (element-wise) product [17] of θ_i^t and θ_j^t to quantify the distribution of their collaboration interests:

$$\boldsymbol{\theta}_{ij}^t = \frac{\boldsymbol{\theta}_i^t \odot \boldsymbol{\theta}_j^t}{\sum_z \left(\boldsymbol{\theta}_i^t \odot \boldsymbol{\theta}_j^t\right)_z} \tag{6}$$

Obviously, the topic $\theta_{ij,z}^t$ has a relatively large value only if both the topics of $\theta_{i,z}^t$ and $\theta_{j,z}^t$ are large. Intuitively, if we have observed that both u_i and u_j have intensive collaboration interest on the topic z during t, i.e. with high proportions $\theta_{i,z}^t$ and $\theta_{j,z}^t$, then z can be regarded as a potential collaboration topic between u_i and u_j.

B. PROBLEM FORMULATION

As presented above, social relationships and collaboration interests are two major reasons leading to changes in coauthorships. As a common principle stated by the *structural balance theory* (a.k.a. "the friend of my friend is my friend"), a researcher may tend to develop new coauthorships with his/her indirect coauthors. Likewise, the shift of interests also influences the formation of new collaborations and vice versa. Such coevolution of both social relationships and research interest has a significant impact on future collaboration, as shown in Figure 1.

According to above analysis, obviously finding suitable coauthors in different years will have different answers, so we need to formulate this problem in a temporally dependent manner. From the probabilistic model view, we can model the probability of each output, y_{ij}^t, at time t conditional to its historical input features. Here, $y_{ij}^t \in \{1, 0\}$ is a Bernoulli variable to indicate whether u_i and u_j have coauthored papers at time t given the historical features of both social relationships $\left\{\mathbf{s}_{ij}^t\right\}$ and the collaboration interests $\left\{\boldsymbol{\theta}_{ij}^t\right\}$. Formally, we can obtain the following conditional probability of the collaboration between u_i and u_j:

$$p\left(y_{ij}^t|\mathbf{s}_{ij}^{t-1}, \cdots, \mathbf{s}_{ij}^{t-N}, \boldsymbol{\theta}_{ij}^{t-1}, \cdots, \boldsymbol{\theta}_{ij}^{t-M}; \boldsymbol{\beta}\right) \tag{7}$$

where the historical social features are modeled of time-slice order N and the historical collaboration topics are modeled of time-slice order M. $\boldsymbol{\beta}$ are model parameters that need to be learned. To simplify discussion, we will assume $N = M$ and refer to N as the time-slice order of the model.

Given a target user u_i, we can write the joint probability of all possible collaboration with u_i from time $t = 1 \cdots T$:

$$P_i = p\left(\boldsymbol{\beta}|\boldsymbol{\lambda}\right) \prod_{t=1}^{T} \prod_{j\neq i} p\left(y_{ij}^t|\mathbf{s}_{ij}^t, \boldsymbol{\theta}_{ij}^t; \boldsymbol{\beta}\right) \tag{8}$$

where $p\left(\boldsymbol{\beta}|\boldsymbol{\lambda}\right)$ is the prior of the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are hyper parameters. The specific distributions of $p\left(\boldsymbol{\beta}|\boldsymbol{\lambda}\right)$ will be discussed in detail in the next section. Therefore, we can learn $\boldsymbol{\beta}$ by maximizing the joint probability P_i.

When the model parameters $\boldsymbol{\beta}$ were estimated, the probability of u_i choosing u_j for coauthoring at time $T + 1$ can be predicted using the same form as Eq. 7:

$$h_{ij}^{T+1} = p\left(y_{ij}^{T+1}|\mathbf{s}_{ij}^{T}, \cdots, \mathbf{s}_{ij}^{T+1-N}, \boldsymbol{\theta}_{ij}^{T}, \cdots, \boldsymbol{\theta}_{ij}^{T+1-M}; \boldsymbol{\beta}\right) \tag{9}$$

As a result, we can compute the probability over all candidates $\mathcal{C}_i$, $\left\{h_{ij}^{T+1}|j \in \mathcal{C}_i\right\}$. Then, we can rank the future potential coauthors of u_i by sorting $\left\{h_{ij}^{T+1}|j \in \mathcal{C}_i\right\}$.

IV. MODEL SPECIFICATION

As discussed above, the formation of collaboration between each pair of researchers is not arbitrary; a researcher is more likely to choose a coauthor whose collaboration can produce more outputs, e.g. publications. In this section, we formally study researchers' choices of their coauthors which are dependent on the time-varying features, i.e. social relationships and collaboration interests.

Choice analysis attempts to model the decision process of an individual in a particular context. Specially, discrete choice models [4], [18] describe, explain, and predict choices between two or more discrete alternatives. In this section, we design a time-aware paired choice model to analyze the choices of coauthors based on utility theory, where we compare the utility of a pair of choices between two candidate coauthors within some time period. Therefore, the problem of coauthorships prediction is reduced to measure the utility of each choice based on the time-varying features.

A. PAIRED CHOICE ANALYSIS

In general, discrete choice models are usually derived from utility theory [5]. Utility is a representation of preferences over a set of alternatives. Given a target researcher u, we represent the preference relation on his/her candidate coauthors, using the notation $\succ_u$. The candidate coauthor set of u is denoted as $\mathcal{C}_u \subseteq \mathcal{U}$. For example, if u prefers to collaborate with a coauthor i over j where $i, j \in \mathcal{C}_u$, then we can formally write such preference as $i \succ_u j$. Furthermore, we define the utility function $U(u, i)$ to quantify the utility of u choosing a researcher i as the coauthor. Then, we have the necessary and sufficient condition as Eq. 10 according to utility theory. Preferences have the utility representation so long as they are transitive, complete, and continuous [5].

$$i \succ_u j \Leftrightarrow U(u, i) > U(u, j) \tag{10}$$

1) ORDER OF PREFERENCE

Intuitively, more outputs (i.e. coauthored papers) mean that the collaboration is more productive, so the preference of collaboration can be measured by the number of coauthored papers between two researchers. Some researchers may have conducted collaboration, but the observed output is zero. Obviously, such collaboration is not preferred. So far, we can define the preference relation at time t according to the number of observed collaborations as follows:

$$i \succ_u^t j \Leftrightarrow \#col_{ui}^t > \#col_{uj}^t \tag{11}$$

If $\#col_{ui}^t = \#col_{uj}^t$, we cannot immediately tell which coauthor is preferred by u, so the preference order is undefined in such a case.

Obviously, the $\succ_u^t$ is a partial order relation over $\mathcal{C}_u$ according to the above definition. Figure 2 visualizes the Hasse diagram of the poset $\mathcal{C}_u$, where the higher position of coauthors indicates the higher number of observed coauthoring, that is, u prefers collaborating with researchers in higher positions than those in lower positions. Furthermore, we defined a set of coauthors for whom u is less preferred than i (note the preference relation is transitive), and we formally denote this set as $\prec_u^t i$:

$$\prec_u^t i = \left\{j|i \succ_u^t j, j \neq i\right\} \tag{12}$$

For those coauthors at the bottom level in Figure 2, they are the least preferred by u with zero coauthoring. Hence, we can define these coauthors as:

$$0_u^t = \left\{j| \prec_u^t j = \varnothing\right\} \tag{13}$$

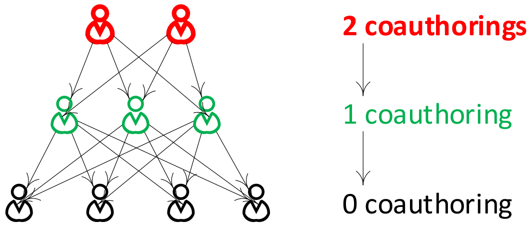


FIGURE 2. The Hasse diagram of the poset $\mathcal{C}_u$ w.r.t. a target researcher u. The order relation $\succ_u^t$ is organized according to the number of observed coauthoring during t . For example, the demonstration illustrates u coauthored with the top-level authors twice and the bottom-level authors with zero time.

And those coauthors who have at least one coauthoring during t are denoted as:

$$\bar{\mathcal{C}}_u^t = \mathcal{C}_u \setminus 0_u^t \tag{14}$$

2) PAIRWISE CHOICE MODELING

The method of pairwise comparison has been widely used in the scientific study of preferences, which can be dated back to the Law of Comparative Judgment that was conceived by Thurstone [19] in 1927. The essential idea behind Thurstone's process and model are that it can be used to scale preference over each choice object based on simple comparisons between a pair of choice objects a time: that is, based on a series of pairwise comparisons.

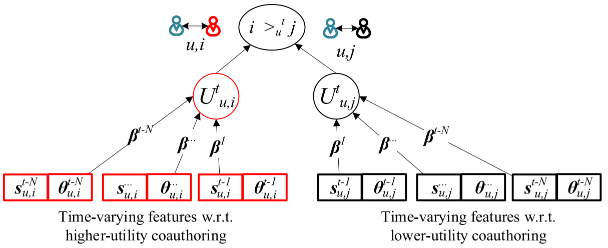


FIGURE 3. Demonstration of the TAPCA model: it models the utilities of user *u* choosing a pair of coauthors *i* and *j* at time *t*, where the utility of coauthering between *u* and *j*, i.e. $U^t_{u,i}$, is higher than that of between *u* and *j*, i.e. $U^t_{u,j}$. $i \succ^t_u j$ denotes the order of pairwise comparison between $U^t_{u,i}$ and $U^t_{u,j}$.

Figure 3 illustrates the model of TAPCA by comparing the utility of choosing a pair of coauthors, the details are presented below. Following the derivation of choice models using the random utility model (RUM) [4], we decompose the utility $U^t_{u,i}$ at time t into two parts:

$$U^t_{u,i} = V\left(\mathbf{x}^t_{ui}\right) + \varepsilon^t_i \tag{15}$$

where the function $V\left(\mathbf{x}^t_{ui}\right)$ is often called *representative utility* and error term ε^t_i captures the factors that affect utility but are not included in $V\left(\mathbf{x}^t_{ui}\right)$. $\mathbf{x}^t_{ui}$ is the input features of the choice i by u at time t. In our problem, the observed features are social features and collaborative topics, namely $\mathbf{s}^t_{ui}$ and $\boldsymbol{\theta}^t_{ui}$. Here, $\boldsymbol{\theta}^t_{ui}$ uses the definition in Eq. 6. $V\left(\mathbf{x}^t_{ui}\right)$ has the linear form as follows:

$$V\left(\mathbf{x}^t_{ui}\right) = \beta_0 + \sum_{n=1}^{N} \boldsymbol{\beta}_s^{n\top} \mathbf{s}^{t-n}_{ui} + \boldsymbol{\beta}_\theta^{t\top} \boldsymbol{\theta}^{t-n}_{ui} \tag{16}$$

where N is the order of this time series model, β_0 is a bias, $\boldsymbol{\beta}^n_s$ is a vector of effects of social features and $\boldsymbol{\beta}^n_\theta$ is a vector of effects of collaboration topics. That is, $\boldsymbol{\beta}^n_s$ models which social features have heavy impact on the choice of coauthors while $\boldsymbol{\beta}^n_\theta$ models a researcher's preference to collaborate on which topics. We can denote $\mathbf{a}^{t\top}_{ui} = \left[\mathbf{s}^{t\top}_{ui}, \boldsymbol{\theta}^{t\top}_{ui}\right]$, $\mathbf{x}^{t\top}_{ui} = \left[\mathbf{a}^{t-n\top}_{ui}\right]^N_{n=1}$ to stack all N-order input features, and $\boldsymbol{\beta}^{n\top} = \left[\boldsymbol{\beta}^{n\top}_s, \boldsymbol{\beta}^{n\top}_\theta\right]$, $\boldsymbol{\beta}^\top = \left[\boldsymbol{\beta}^{n\top}\right]^N_{n=1}$ to stack all corresponding model parameters, and then Eq. 16 can be rewritten in a very concise form:

$$V\left(\mathbf{x}^t_{ui}\right) = \beta_0 + \sum_{n=1}^{N} \boldsymbol{\beta}^{n\top} a^{t-n}_{ui} = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}^t_{ui} \tag{17}$$

According to Thurstone's Law of Comparative Judgment Case 5 [19], the probability of preference can be defined in terms of the utility of choices (Eq. 10), namely

$$p\left(i \succ^t_u j\right) = p\left(U^t_{u,i} > U^t_{u,j}\right) \tag{18}$$

Further, we replace $U^t\,(u, j)$ with Eq. 15, and then we obtain:

$$\begin{aligned} &p\left(U^t_{u,i} > U^t_{u,j}\right) \\ &\quad = p\left(V\left(\mathbf{x}^t_{ui}\right) + \varepsilon^t_i > V\left(\mathbf{x}^t_{uj}\right) + \varepsilon^t_j\right) \\ &\quad = p\left(\varepsilon^t_j < \varepsilon^t_i + V\left(\mathbf{x}^t_{ui}\right) - V\left(\mathbf{x}^t_{uj}\right)\right) \\ &\quad = CDF\left(\varepsilon^t_i + \boldsymbol{\beta}^\top \mathbf{x}^t_{ui} - \boldsymbol{\beta}^\top \mathbf{x}^t_{uj}\right) \end{aligned} \tag{19}$$

where CDF denotes some cumulative density function.

If we assume $\varepsilon^t_i \sim extremevalue$, the logit model is obtained by integrating out ε^t_i from Eq. 19 with a closed form [4]. Finally, the probability of preference over a pair of choices is represented by a logistic function as given by Eq. 20.

$$\begin{aligned} p\left(i \succ^t_u j | \boldsymbol{\beta}\right) &= \frac{e^{V\left(\mathbf{x}^t_{ui}\right)}}{e^{V\left(\mathbf{x}^t_{ui}\right)} + e^{V\left(\mathbf{x}^t_{uj}\right)}} \\ &= \frac{1}{1 + e^{-\boldsymbol{\beta}^\top\left(\mathbf{x}^t_{ui} - \mathbf{x}^t_{ui}\right)}} \end{aligned} \tag{20}$$

Further, we can conduct pairwise comparisons over all possible preferences, $\succ^t_u$, derived from the poset $\mathcal{C}_u$. As a result, the joint likelihood over all $\succ^t_u$ is given by:

$$p\left(\succ^t_u | \boldsymbol{\beta}\right) = \prod_{i \in \bar{\mathcal{C}}^t_u} \prod_{j \in \prec^t_u i} p\left(i \succ^t_u j | \boldsymbol{\beta}\right) \tag{21}$$

Furthermore, we jointly consider the time series of input features and output collaborations from $t = 1$ to T. The final likelihood is obtained as follows:

$$p\left(\succ_u | \boldsymbol{\beta}\right) = \prod_{t=N+1}^{T} p\left(\succ^t_u | \boldsymbol{\beta}\right) \tag{22}$$

B. PRIORS ON PARAMETERS

So far, we have presented the likelihood function, so we can use MLE (Maximum Likelihood Estimation) to estimate the model parameters $\boldsymbol{\beta}$ from Eq. 22. However, such estimates of $\boldsymbol{\beta}$ lack regularization, which may lead to poor prediction performance due to overfitting. Therefore, we can complete our model as Bayesian modeling by placing some prior $p\left(\boldsymbol{\beta} | \boldsymbol{\Theta}\right)$ on the model parameters $\boldsymbol{\beta}$, where $\boldsymbol{\Theta}$ denotes hyperparameters. In this paper, we consider two frequently used priors:

1) GAUSSIAN PRIOR

First, we place a spherical multivariate Gaussian prior on the parameter vector $\boldsymbol{\beta}$ with zero means and a diagonal covariance matrix specified by $\sigma^2\mathbf{I}$:

$$p\left(\boldsymbol{\beta}|\boldsymbol{\Theta}\right) = N\left(\boldsymbol{\beta}|0, \sigma^2\mathbf{I}\right) = \lambda_0 e^{-\frac{\sum_{k=1}^{K}\beta_k^2}{2\sigma^2}} \tag{23}$$

where σ^2 is a variance parameter, $\mathbf{I}$ is the identity matrix and λ_0 is the normalization constant.

2) LAPLACE PRIOR

Second, we consider the Laplace prior with zero means and variance parameter σ^2 for each β_k, so the prior density for the parameter vector $\boldsymbol{\beta}$ is

$$p\left(\boldsymbol{\beta}|\boldsymbol{\Theta}\right) = \prod_{k=1}^{K} LAP\left(\beta_k|0, \sigma^2\right) = \lambda_0 e^{-\sqrt{2}\frac{\sum_{k=1}^{K}|\beta_k|}{\sigma^2}} \tag{24}$$

where K is the length of β and λ_0 is the normalization constant.

C. OBJECTIVE FUNCTION

With the likelihood (Eq. 22.) and priors (Eq. 23 and Eq. 24) in hand, we immediately obtain the posterior by Bayes theorem.

$$p\left(\boldsymbol{\beta}|\succ_u\right) \propto p\left(\succ_u|\boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}|\boldsymbol{\Theta}\right) \tag{25}$$

Therefore, we can maximize the posterior to learn β Obviously, maximizing Eq. 25 is equivalent to minimizing its negative log form. Then, the loss function is obtained:

$$L_{\boldsymbol{\beta}}\left(\succ_u\right) = \arg\min_{\boldsymbol{\beta}} -\log p\left(\succ_u|\boldsymbol{\beta}\right) - \log p\left(\boldsymbol{\beta}|\boldsymbol{\Theta}\right) \tag{26}$$

In the above loss function, the term $R\left(\boldsymbol{\beta}\right) = -\log p\left(\boldsymbol{\beta}|\boldsymbol{\Theta}\right)$ servers as the regularizer to avoid overfitting.

For the case of Gaussian prior (Eq. 23), the regularizer $R_N\left(\boldsymbol{\beta}\right)$ corresponds to the L2-norm regularization:

$$R_N\left(\boldsymbol{\beta}\right) = -\log N(\boldsymbol{\beta}|0, \sigma^2\mathbf{I}) = \lambda\|\boldsymbol{\beta}\|_2^2 = \lambda\sum_{k=1}^{K}\beta_k^2 \tag{27}$$

where $\lambda > 0$ is the regularization parameter which can be tuned by cross-validation. The L2-norm regularizer plays a role in shrinking $\boldsymbol{\beta}$ to relatively small values.

For the case of Laplace prior 24, the regularizer $R_L\left(\boldsymbol{\beta}\right)$ corresponds to the L1-norm regularization:

$$R_L\left(\boldsymbol{\beta}\right) = -\log LAP(\boldsymbol{\beta}|0, \sigma^2\mathbf{I}) = \lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{k=1}^{K}|\beta_k| \tag{28}$$

The Laplace prior has thinner tails than the Gaussian, and thus concentrates posteriors closer to its zero mean than a Gaussian of the same variance. Therefore, the parameters β with L1-norm regularizer tend to induce sparse values, i.e. only a few $\boldsymbol{\beta}_k$ are non-zero and the other ones are zeroes.

V. LEARNING AND INFERENCE

As presented in previous sections, we need to learn the parameters $\boldsymbol{\beta}$ of the time-series input features in order to find future coauthors. In this section, we design a gradient-based algorithm to efficiently learn the parameters, and then we show how to prediction future coauthorships.

A. PARAMETER ESTIMATION

Minimizing loss function $L_{\boldsymbol{\beta}}\left(\succ_u\right)$ (Eq. 26) can be viewed as solving a regularized logistic regression problem. In general, we can use a gradient-based method to find $\boldsymbol{\beta}$. The gradient of $L_{\boldsymbol{\beta}}\left(\succ_u\right)$ can be computed as the follows.

$$\nabla L_{\boldsymbol{\beta}}\left(\succ_u\right) = -\frac{\partial \log p\left(\succ_u|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} - \frac{\partial \log p\left(\boldsymbol{\beta}|\Theta\right)}{\partial\boldsymbol{\beta}} \tag{29}$$

First, the partial derivatives of the likelihood function can be easily derived from Eq. 20-22:

$$\frac{\partial \log p\left(\succ_u|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} = \sum_{t=N+1}^{T}\sum_{i\in\bar{\mathcal{C}}_u^t}\sum_{j\in\prec_u^t i}\frac{\partial \log p\left(i \succ_u^t j|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} \tag{30}$$

where

$$\frac{\partial \log p\left(i \succ_u^t j|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} = \frac{-e^{-\boldsymbol{\beta}^\top(\mathbf{x}_{ui}^t-\mathbf{x}_{ui}^t)}}{1+e^{-\boldsymbol{\beta}^\top(\mathbf{x}_{ui}-\mathbf{x}_{uj})}}\mathbf{x}_{ui}^t \tag{31}$$

Moreover, we can give the partial derivatives of the regularization w.r.t. L2 and L1 norms.

1) *L2 Regularization* (cf. Eq. 27):

$$-\frac{\partial \log p\left(\boldsymbol{\beta}|\Theta\right)}{\partial\boldsymbol{\beta}} = \frac{\partial R_N\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} = \lambda\sum_{k=1}^{K}\beta_k \tag{32}$$

2) *L1 Regularization* (cf. Eq. 28):

$$-\frac{\partial \log p\left(\boldsymbol{\beta}|\Theta\right)}{\partial\boldsymbol{\beta}} = \frac{\partial R_L\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} = \lambda\sum_{k=1}^{K}\mathrm{sgn}(\beta_k) \tag{33}$$

The number of possible paired choices, $i \succ_u^t j$, is large, which can be estimated using Eq. 34 where $|\cdot|$ denotes the cardinality.

$$|\succ_u^t| \approx |\bar{\mathcal{C}}_u^t| * |\mathcal{C}_u| \tag{34}$$

The coauthor candidate set $\mathcal{C}_u$ often consists of all available researchers, i.e. $\mathcal{U}$. As a result, $|\succ_u^t|$ tends to be large. As a result, standard gradient descent methods are not workable for our problem. To solve this issue, we use the gradient-based optimization algorithm to estimate the parameters. Furthermore, we adopt the bootstrap sampling strategy to randomly draw a paired choice when running our learning algorithms. Such a strategy has been proved very efficient in large-scale problems [20].

1) LEARNING ALGORITHM

Given a training example, i.e. a pair of choices, $i \succ_u^t j$ at time t, the gradient is given by:

$$\nabla L_{\boldsymbol{\beta}}\left(i \succ_u^t j\right) = -\frac{\partial \log p\left(i \succ_u^t j|\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} + \frac{\partial R\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} \tag{35}$$

In fact, we do not need to loop the stochastic gradient update for all possible choice pairs. In this work, we take a bootstrap sampling strategy to randomly draw a batch of training samples $\mathcal{B} = \left\{i \succ_u^t j\right\}$, which has been proved very efficient and quick converging over large-scale data [20]. Accordingly, we design the learning scheme with L1 and L2 regularization given by Algorithm 1, where η denotes the learning rate that is adaptively optimized in terms of Adam [21].

Algorithm 1 Gradient-Based Learning Scheme for TAPCA

```
Require:
  $\mathcal{U}$: the author set
  $\lambda$: regularization parameter
  $\eta_0$: initial learning rate
  $N$: time-slice order of the time series model
  $B$: mini-batch size
  $c$: sampling factor
  $MaxIt$: maximum number of iterations
Ensure: $\boldsymbol{\beta}$ the learned parameters w.r.t. $u$
  $\boldsymbol{\beta} \leftarrow 0$, $\eta \leftarrow \eta_0$
  $K \leftarrow c \sum_{t=N+1}^{T} \sum_{u\in\mathcal{U}} |\bar{\mathcal{C}}_u^t|/B$        ▷ number of batches
  for $it = 1$ to $MaxIt$ do
      for $k = 1$ to $K$ do
          for $b = 1$ to $B$ do
              sample a target researcher $u$ from $\mathcal{U}$
              sample a coauthor $i$ from $\mathcal{C}_u \setminus \mathbf{0}_u^t$
              sample a less preferred coauthor $j$ from $\prec_u^t i$
              add $i \succ_u^t j$ to training batch $\mathcal{B}$
          end for
          $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \eta \frac{\sum_{i\succ_u^t j\in\mathcal{B}}(\nabla L_{\boldsymbol{\beta}}(i\succ_u^t j))}{B}$        ▷ cf. Eq. 35
          $\eta \leftarrow Adam(\eta)$
      end for
  end for
  return $\boldsymbol{\beta}$
```

2) COMPLEXITY ANALYSIS

According to Eq. 34, the number of all possible training samples is $\sum_{t=N+1}^{T} \sum_{u\in\mathcal{U}} | \ \succ_u^t \ |$, with the complexity $O(|\mathcal{U}| \sum_{t=N+1}^{T} \sum_{u\in\mathcal{U}} |\bar{\mathcal{C}}_u^t|)$ for each iteration, which cannot be efficiently computed on large-scale datasets. When we adopt the bootstrap sampling strategy as presented in Algorithm 1, the complexity of each iteration is reduced to $O\left(c \sum_{t=N+1}^{T} \sum_{u\in\mathcal{U}} |\bar{\mathcal{C}}_u^t|\right)$. In general, the sampling factor $c \ll |\mathcal{U}|$. In this paper, we set the contrastive pair factor $c = 10$, that is, for each observed collaboration we only sample 10 less preferred coauthor for pairwise comparisons. In this way, our algorithm can be finished in constant time for each iteration.

B. COAUTHORSHIP PREDICTION

According to utility theory, researchers' choices over coauthors are measured by utility (cf. Eq. 10). The higher utility implies the more preferred coauthor. Therefore, we can rank future coauthors in terms of the predictive utility. Given a target researcher u, the expectation of the predictive utility of choosing i as a coauthor at the future time $t = T + 1$ (cf. Eq. 34) can be computed by the expectation of Eq. 15 with the estimated model parameters $\boldsymbol{\beta}$, where $const = \beta_0 + \mathbb{E}\left(\varepsilon_i^t\right)$ is a constant. In the same way, we can compute utility w.r.t. every potential coauthor from the candidate set C_u.

$$\begin{aligned} \mathbb{E}\left[U_{u,i}^t|\boldsymbol{\beta}\right] &= \mathbb{E}\left[V\left(\mathbf{x}_{ui}^t\right) + \varepsilon_i^t|\boldsymbol{\beta}\right] \\ &= V\left(\mathbf{x}_{ui}^t\right) + \mathbb{E}\left(\varepsilon_i^t\right) \\ &= \boldsymbol{\beta}^{\top}\mathbf{x}_{ui}^t + const \end{aligned} \tag{36}$$

Then, we can simply rank the future coauthors according to their *representative utility* $\left\{\boldsymbol{\beta}^{\top}\mathbf{x}_{ui}^t\right\}_{i\in C_u}$.

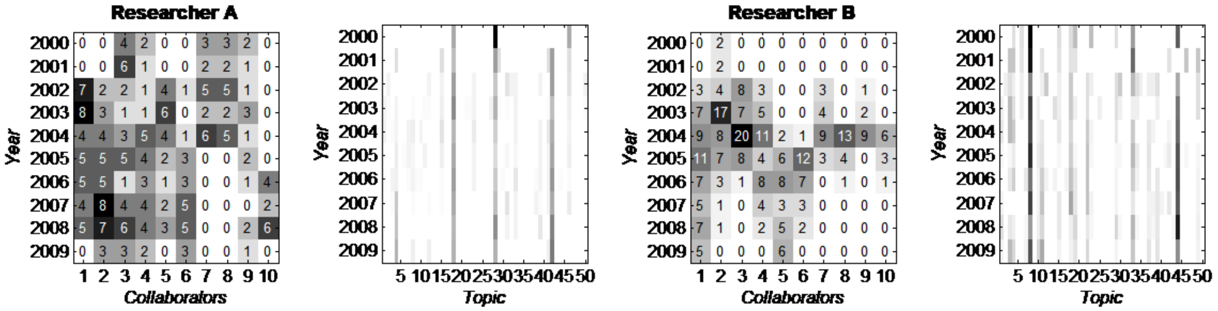
VI. EXPERIMENTS

The experiments were conducted on a real-world academic dataset. This dataset was collected using the Microsoft Academic Search (MAS) API from its academic database. We evaluate the prediction performance by a set of metrics and compared our models with other state-of-the-art approaches. We implemented our model using Keras [22] with the backend of Tensorflow GPU version. Gensim [23] is employed to extract topics from the corpus. The experiments were run on a machine with 32G memory and 8G GPU.

A. DATA PREPARATION

First, we selected eighteen computer science conferences: SIGIR, WWW, KDD, IJCAI, AAAI, ICML, ICDE, CIKM, ICDM, VLDB, CVPR, SIGMOD, NIPS, ACL, UAI, PKDD, SDM, and PAKDD, which cover the research areas: information retrieval, machine learning, artificial intelligence, database and data mining. We respectively retrieved the top 200 researchers with the maximum numbers of publications from each conference using the MAS API. Then, these researchers were merged into a researcher set, resulting in a total of 2,137 distinct researchers. Finally, we used the MAS API to retrieve the articles published from 2000 to 2010 w.r.t. each researcher and a total of 76,431 articles were collected.

We extracted the words from the abstract and title for each article and removed stop words and words appearing less than five times. This preprocessing yielded a vocabulary containing 8,700 distinct words. Since this dataset focuses on the publications of AI-related area, we select 50 latent topics for the LDA model to retrieve the empirical topic distribution for each researcher for each year (cf. Eq. 5). For some multi-discipline datasets, we may choose a larger number of latent topics, e.g., 300, to study the coauthoring on cross-discipline topics. Moreover, we constructed a set of coauthorship networks over those 2,137 researchers according to their coauthorships retrieved from the articles, where each network corresponds to a year. In addition, we also



Researcher A

Year	1	2	3	4	5	6	7	8	9	10
2000	0	0	4	2	0	0	3	3	2	0
2001	0	0	6	1	0	0	2	2	1	0
2002	7	2	2	1	4	1	5	5	1	0
2003	8	3	1	1	6	0	2	2	3	0
2004	4	4	3	5	4	1	6	5	1	0
2005	5	5	5	4	2	3	0	0	2	0
2006	5	5	1	3	1	3	0	0	1	4
2007	4	8	4	4	2	5	0	0	0	2
2008	5	7	6	4	3	5	0	0	2	6
2009	0	3	3	2	0	3	0	0	1	0

Researcher B

Year	1	2	3	4	5	6	7	8	9	10
2000	0	2	0	0	0	0	0	0	0	0
2001	0	2	0	0	0	0	0	0	0	0
2002	3	4	8	3	0	0	3	0	1	0
2003	7	17	7	5	0	0	4	0	2	0
2004	9	8	20	11	2	1	9	13	9	6
2005	11	7	8	4	6	12	3	4	0	3
2006	7	3	1	8	8	7	0	1	0	1
2007	5	1	0	4	3	3	0	0	0	0
2008	7	1	0	2	5	2	0	0	0	0
2009	5	0	0	0	6	0	0	0	0	0

FIGURE 4. **The statistics of the two most active researchers: (Left) the number of coauthored papers with their top 10 collaborators each year; (Right) the distribution of 50 collaboration topics each year (the grey degree indicates the research interest intensity on one specific topic).**

constructed a single coauthorship network over the training data, ignoring the time for baseline models in a static manner. In the following experiments, the data of years 2000-2009 are used for training, and the remaining data of the year 2010, i.e. the true coauthorships in 2010, are used for testing. The candidate coauthor set $\mathcal{C}_u$ for each author consists of all researchers in the dataset.

Figure 4 visualizes coauthorships and collaboration topics w.r.t. two active researchers over ten years. From this demonstration, we can clearly observe the evolution of coauthorships and research focus over time.

B. EVALUATION METRICS

Since coauthorship prediction is a process of ranking, we use the following three metrics, which are commonly used in IR and search to evaluate the performance of all comparison models. In the following experiments, we reported the average performance over all testing users.

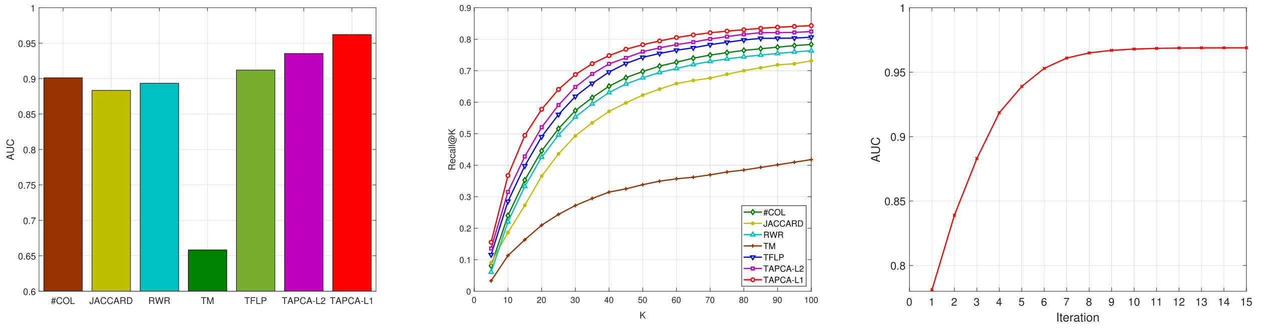
- *AUC:* Area under the ROC curve measures the probability that a system ranks a positive instance higher than a negative one. Given a target researcher u, AUC_u evaluates the probabilities of correctly assigning the ranks for the true observed coauthors cl_u (positive instances) higher than the remaining $\mathcal{C}_u \setminus cl_u$ (negative ones):
$$AUC_u = \frac{\sum_{i \in cl_u} \sum_{j \in \mathcal{C}_u \setminus cl_u} 1\left(rk\left(i\right) \leq rk\left(j\right)\right)}{|cl_u| \cdot |\mathcal{C}_u \setminus cl_u|} \tag{37}$$
where $rk\left(i\right)$ is a function to retrieve the rank of coauthor i, $1\left(rk\left(i\right) \leq rk\left(j\right)\right)$ is a indicate function which returns 1 if the rank of coauthor i in cl_u is higher than that of coauthor j not in cl_u, and returns 0 otherwise. Note that $rk\left(i\right)$ returns the rank number, i.e. the smaller the value, the higher the rank.
- *Recall@k:* the recall of top k predicted coauthors for a target researcher u is defined by:
$$Recall_u@k = \frac{|\widehat{cl}_u@k \cap cl_u|}{|cl_u|} \tag{38}$$
where $\widehat{cl}_u@k$ denotes the top k predicted coauthors for u and cl_u is the true coauthors in the testing set.
- *nDCG@K:* Normalized Discounted Cumulative Gain [24] is a metric sensitive to the prediction order, so it is used to evaluate the performance of ranking algorithms. *nDCG* assigns a different relevance score to each retrieved item. The highly relevant items appearing lower in the predicted ranking list should be penalized.
$$nDCG_u@k = \frac{DCG_u@k}{IDCG_u@k} \tag{39}$$
$$DCG_u@k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{log_2 i} \tag{40}$$

The above equations give the definition of $nDCG_u@k$ for each researcher u, where rel_i is the graded relevance of the result at position i and $IDCG_u@k$ refers to the $DCG_u@k$ with an ideal ordering. We use binary relevance values: $rel_i \in \{1, 0\}$.
$$rel_i = \begin{cases} 1 & rk\left(cl_i\right) \leq i \\ 0 & otherwise \end{cases} \tag{41}$$
where $rk\left(cl_i\right)$ is the true rank of the coauthor cl_i with the predicted rank i. Such a setting can be interpreted that if the true rank is higher than the position i, then the relevance is 1 and 0 otherwise.

C. COMPARISON METHODS

In all the following experiments, a group of state-of-the-art methods are evaluated for comparison, including our models.

1) *#COL:* This simply ranks predicted coauthors according to the total number of coauthoring.
2) *JACCARD:* This ranks predicted coauthors according to Jaccard similarity as used by CollabSeer [2], where $\Gamma\left(u\right)$ returns all neighbor nodes of u on a coauthorship network.
$$VS_{Jaccard}\left(u_i, u_j\right) = \frac{|\Gamma\left(u_i\right) \cap \Gamma\left(u_j\right)|}{|\Gamma\left(u_i\right) \cup \Gamma\left(u_j\right)|} \tag{42}$$
3) *RWR:* This ranks coauthors by running random walk with restart on the coauthorship network [13].
4) *TM:* This ranks coauthors by running RWR where the transition probability is proportional to the similarity of topic distributions between two researchers. A similar comparison method is used in CTL [3] for matching the topic proportions of two authors in different domains.



(a) AUC

(b) Recall@5∼100

(c) Converging status with increasing iteration

FIGURE 5. The AUC and Recall performance of all comparison models and the convergence test.

Here, we directly match the topic distributions without the setting of domains.

5) *TFLP:* This uses a temporal link prediction method [14], [15] to predict future coauthors by tensor factorization. Each slice of the tensor corresponds to a coauthorship network of a year, i.e. the entry (i,j,t) is 1 if the researchers i and j coauthored papers in the year t, otherwise this entry is 0.

6) *TAPCA-L1:* L1 regularized TAPCA model. This ranks coauthors in terms of utility (cf. Eq. 36). In the following experiments, we set the time-slice order $N = 3$ for the model and hyper parameter $b = 2$ for the long-term social features (cf. Table 1) if we do not explicitly specify other settings. For the learning algorithm, we set the initial learning rate $\eta_0 = 0.001$, the regularization parameter $\lambda = 0.0001$ and the batch size $B = 200$.

7) *TAPCA-L2:* L2 regularized TAPCA model. In the experiments, the setting of hyperparameters is the same as *TAPCA-L1*.

8) *TAPCA-S:* For ablation test, we only use the time-varying social feature (cf. Eq. 1) without involving collaboration topics.

D. EVALUATION RESULTS

1) TARGETING FUTURE COAUTHORS

In general, we can measure the quality of prediction by testing whether the target researcher is more likely to conduct coauthoring with the predictive ones according to the observed coauthorships in real cases. Therefore, it can be viewed as a link prediction problem which is mostly evaluated with AUC.

We evaluated the prediction result for each researcher and report the average AUC over all researchers. The results of comparison approaches are illustrated in Figure 5(a). From the results, we can see that the TM model significantly underperforms other models. This is because collaboration is often based on some common interests, but the inverse is not true. For example, a researcher may have common interests with a distinguished scholar. Obviously, it cannot be concluded that the distinguished scholar will collaborate with the researcher. Therefore, it is not practical to predict coauthorships only in term of matching their research interests. In comparison, other models achieve much better performance than TM. Here the major difference is that all methods, except for the TM model, include the feature of social relationships. Hence, we can conclude that the formation of collaboration is heavily dependent on social relationships.

Moreover, TAPCA-S achieves better performance than #COL, JACCARD and RWR models, which illustrates that TAPCA-S can more effectively capture the shift of social factors over time to better predict the trends in the near future. Furthermore, although both TAPCA-S and TFLP consider the temporal shift, TAPCA-S outperforms TFLP by a significant margin. This is because TFLP shallowly models the change of coauthorships over years; instead, TAPCA-S more effectively both short-term social impact and long-term social impact (cf. Eq. 1). In comparison, TAPCA-L1 and TAPCA-L2 achieve better performance than TAPCA-S. This is because TAPCA-L1 and TAPCA-L2 additionally model the collaboration research topics which are one of the driven forces leading to coauthorings. In particular, TAPCA-L1 outperforms TAPCA-L2 due to the sparse representation. Intuitively, only a few factors are decisive when a researcher chooses their coauthors, which is closer to the choice of coauthors in a real situation. For example, a researcher i may choose j as a coauthor only because j is the friend of most of i's friends (i.e. heavily determined by the feature $\#coc_{ij}$, cf. Table 1). Moreover, a researcher often focuses on very few topics for collaboration instead of all topics, that is, not all the collaboration topics need to be involved. Therefore, the prediction results using sparse representation induced by L1 regularization achieve the best performance on this dataset.

For a real-world searching system, the size of an effective retrieval set is small, normally less than 20, because users are often only patient in browsing the results on the first page. Hence we use *Recall@K* to measure the number of successfully predicted cases within the top K ranked coauthors. Figure 5(b) depicts the average results of Recall@5∼ 100 over all researchers. We find that the plots of TAPCA models are higher than those of other models by a margin. In particular, we find the margin is relatively large for $K \leq 20$ so TAPCA models return a much better result set that contains more effective coauthors than all the other approaches. Similar to the evaluation using AUC, the performance of the TM model is very poor because it does not make sense of predicting coauthorships by simply matching topics between researchers. TAPCA-L1 again achieves the best performance.

We further investigate the convergence of our learning algorithms. Figure 5(c) depicts the probing AUC evaluated after each iteration, and we find that our learning algorithms converge very quickly and reach a stable result within 10 generations. Such fast convergence proves the scalability of our algorithms. Since our algorithms can be run in parallel w.r.t. each target user, it is practical to integrate them into a real-world system.

2) RANKING PREFERENCE OVER COAUTHORS

Cooperating with a suitable coauthor often can bring productive results, e.g. a collection of coauthored publications. All researchers hence prefer to find their own suitable coauthors. To evaluate the prediction of ranking on preferred coauthors, we construct the true preference ranking w.r.t. each researcher by sorting the number of coauthored papers with each of coauthors over the testing set.

We evaluated such a ranking problem over all comparison models by the metrics nDCG@5 and nDCG@10. The average result over all researchers is reported in Table 2. The simplest method #COL achieves considerably good performance among all comparison models. Through some further consideration, we find the rationality behind this. The large number of coauthoring between a researcher and his/her coauthor implies their collaboration is very productive. Hence, researchers always prefer to constantly coauthor papers with such suitable coauthors. This can be partially viewed as a "rich get richer″ phenomenon, so #COL is simple but effective.

TABLE 2. **nDCG of comparison models (* for $p < 0.05$).**

Model	nDCG@5	nDCG@10
#COL	0.2062	0.2437
JACCARD	0.1914	0.2227
RWR	0.2025	0.2283
TM	0.0711	0.0835
TFLP	0.2131	0.2396
TAPCA-S	0.2487*	0.2799*
TAPCA-L2	0.2638*	0.2912*
TAPCA-L1	**0.2712***	**0.3015***

However, such static methods have some limitations as time passes. For example, a researcher's best coauthor retired but this coauthor will always have a high rank in subsequent years, because of the large number of historical coauthoring. In comparison, our model considers not only the short term social factors, $\#col_{ij}^{t}$ and $\#coc_{ij}^{t}$, but also the long-term factors, $\#col_{ij}^{[t-b,t]}$ and $\#coc_{ij}^{[t-b,t]}$ (cf. Table 1). Obviously, these social features are not static, and they vary with time t. In the above case, the values of these social features automatically decrease as time goes by, so our time-aware model does not suffer from such an issue. In addition, our model captures a shift in collaboration topics as well as social factors, so the compatibility between two coauthors can be automatically accessed over time. With these advantages, our model obviously can better predict future trends than other comparison methods. Therefore, it is no wonder our TAPCA methods outperform other models by a large margin. We conduct the significance testing on the TAPCA models and the best baseline method, and the results show the statistical significance of TAPCA.

3) SPECIAL CASES STUDY

Our approach relies on historical (consecutive) collaboration dates to make predictions. In the real-world applications, however, such data is sometimes fragmented and incomplete, e.g., (1) some researchers may not publish papers every year; (2) a Ph.D. student has a short publication record with papers only in recent years. How can our model deal with these challenges? Fortunately, we can adjust the hyper-parameters to deal with these two cases, namely changing the time-slice order N for the temporal model (cf. Eq. 22) and the timespan parameter b for long-term features (cf. Table 1 and Eq. 5) across multiple years.

For the first case, the solution is straightforward. For example, a researcher has published papers in the years $\{2001, 2003, 2005, 2007, 2009\}$ instead of every consecutive year. According to Table 1 and Eq. 5, the timespan parameter b controls the statistical granularity of coauthoring. Consequently, we can set $b \geq 2$ to involve the papers published across more than two years.

The second case reflects the fact that some senior researchers may have long historical data whereas some new authors may only have limited data in recent years. This can also be handled by setting the hyper-parameters, i.e., the time-slice order parameter N and the timespan parameter b. For example, we can set $N = 5$ and $b = 3$ for senior researchers with more than ten-year data while $N = 2$ and $b = 1$ for a new author with five-year data. In the extreme case, we can train our model over two-year data, namely, we fit the output data of the second year by regressing on the input feature of the first year (cf. Eq. 22).

In the following experiment, we use the most recent five-year data for training, i.e. 2005∼ 2009, so as to simulate new authors who only have short-term historical data. For TAPCA models, we fixed the time slice order $N = 2$ to test the impact of different timespans $b = \{1, 2, 3\}$. That is, we respectively set the time granularity to be one year, two years and three years to evaluate the performance.

From the results depicted in Table 3, we find the performance is relatively close among all comparison models. This reflects that the research focus and the coauthorships do not dramatically change in the short term. TAPCA models overall outperform baseline models because they consider more comprehensive features. The significance testing also shows that the TAPCA models significantly outperform the other baselines. In particular, we find that the TAPCA models with the setting $b = 2$, i.e., two-year timespan, achieve better performance than the models with the setting $b = 1$ and $b = 3$, which illustrates that the time-varying features of two-year granularity (cf. Table 1 and Eq. 5) best capture the statistics of the coauthoring information according to this dataset. The best setting of b is dependent on specific datasets, it may need to set a smaller b to capture shorter term coauthering information whereas a larger b to capturelonger term coauthoring. To determine the best b, we need to tune it on a validation set.

TABLE 3. AUC and nDCG@10 in terms of different timespans {1,2,3} (* for $p < 0.05$).

Model	AUC	nDCG@10
#COL	0.8276	0.2213
JACCARD	0.8027	0.2130
RWR	0.8223	0.2162
TFLP	0.8279	0.2234
TAPCA-L2 (b=1)	0.8519*	0.2711*
TAPCA-L2 (b=2)	**0.8602***	**0.2803***
TAPCA-L2 (b=3)	0.8587*	0.2782*
TAPCA-L1 (b=1)	0.8576*	0.2775*
TAPCA-L1 (b=2)	**0.8639***	**0.2838***
TAPCA-L1 (b=3)	0.8617*	0.2829*

VII. CONCLUSION AND DISCUSSION

In this paper, we addressed the problem of coauthorship modeling and prediction in a dynamic fashion, which can reflect the nature of evolving coauthorships. Accordingly, we take both time-varying social relationships and collaboration interest into account, and we designed a time-series model to capture the dynamic choices of coauthors over time. In particular, our model is constructed with the idea of time-aware paired choice analysis (TAPCA) to model preferences by pairwise comparisons under the theories of utility and choice modeling. Furthermore, we employ Bayesian modeling to regularize our probabilistic model with various priors so as to improve the generalization. Finally, the experiments evaluated on the real-world academic dataset demonstrate that our approach can more accurately predict future coauthorships than other state-of-the-art methods.

Although we focus on the coauthorship modeling and prediction in this paper, the TAPCA framework is also applicable to analyzing other social networks, especially the online social networking sites, e.g. Twitter, Facebook, and LinkedIn. For example, the repost prediction on Twitter is very similar to coauthorships prediction from the TAPCA view, where users' interested topics change every day and users' active followers are also not static. Therefore, we can easily set up the time-varying features of social relationships and interesting topics (here, the features can be a little different from the features for modeling coauthorships). As a result, we can recommend personalized tweets for users. Take another example, we, in fact, actively or passively develop some new social relationships in our daily life, and our preferred jobs may also change with time. If we apply our TAPCA on the LinkedIn data to model such time-varying features, then we may help users to capture the potential job opportunity.

REFERENCES

[1] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proc. Nat. Acad. Sci. USA*, vol. 101, p. 5200, Apr. 2004.

[2] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles, "CollabSeer: A search engine for collaboration discovery," in *Proc. 11th Annu. Int. ACM/IEEE Joint Conf. Digit. Libraries*, Jun. 2011, pp. 231–240.

[3] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1285–1293.

[4] K. Train, *Discrete Choice Methods with Simulation*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[5] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (60th Anniversary Commemorative Edition). Princeton, NJ, USA: Princeton Univ. Press, 2007.

[6] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, p. 404, 2001.

[7] K. Miller, T. L. Griffiths, and M. I. Jordan, "Nonparametric latent feature models for link prediction," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 1276–1284.

[8] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 635–644.

[9] I. Makarov, O. Bulanov, and L. E. Zhukov, "Co-author recommender system," in *Models, Algorithms, and Technologies for Network Analysis*, V. Kalyagin, A. Nikolaev, P. Pardalos, and O. Prokopyev, Eds. Cham, Switzerland: Springer, 2017, pp. 251–257.

[10] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, 2011, pp. 121–128.

[11] Y. Sun, J. Tang, J. Han, C. Chen, and M. Gupta, "Co-evolution of multi-typed objects in dynamic star networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2942–2955, Dec. 2013.

[12] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Sci.*, vol. 3, no. 1, p. 9, Sep. 2014. doi: 10.1140/epjds/s13688-014-0009-x.

[13] J. Li, F. Xia, W. Wang, Z. Chen, N. Y. Asabere, and H. Jiang, "ACRec: A co-authorship based random walk model for academic collaboration recommendation," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 1209–1214.

[14] E. Acar, D. M. Dunlavy, and T. G. Kolda, "Link prediction on evolving data using matrix and tensor factorizations," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2009, pp. 262–269.

[15] M. Araújo, P. Ribeiro, and C. Faloutsos, "TensorCast: Forecasting time-evolving networks with contextual information," in *Proc. 27th Int. Joint Conf. Artif. Intell., (IJCAI)*, Jul. 2018, pp. 5199–5203. doi: 10.24963/ijcai.2018/721.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[17] J. Chang and D. M. Blei, "Hierarchical relational models for document networks," *Ann. Appl. Statist.*, vol. 4, no. 1, pp. 124–150, 2010.

[18] W. H. Greene and D. A. Hensher, *Modeling Ordered Choices: A Primer*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[19] L. L. Thurstone, "A law of comparative judgment," *Psychol. Rev.*, vol. 34, no. 4, p. 273, 1927.

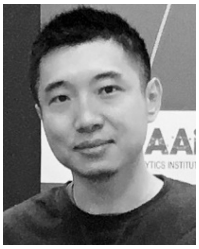
[20] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell. (AUAI)*, 2009, pp. 452–461.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[22] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[23] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50. [Online]. Available: http://is.muni.cz/publication/884893/en

[24] W. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Reading, MA, USA: Addison-Wesley, 2010.



LIANG HU currently holds a postdoctoral position at the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, and a Research Associate with the Advanced Analytics Institute, University of Technology Sydney, Australia. His research interests include recommender systems, data mining, machine learning, representation learning, and general artificial intelligence. He has published a number of papers in top-rank international conferences and journals, including WWW, IJCAI, AAAI, ICDM, and ICWS. He has published a number of papers in top-rank international journals such as TOIS and JWSR.

QINGKUI CHEN is currently a Professor and a Ph.D. Supervisor with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology (USST), Shanghai, China. He is the Vice Dean of the School of Optical Electrical and Computer Engineering. His research interests include network computing, parallel computing, parallel database, and computer networks. He is the head of many programs which were supported by the Natural Science Foundation of China (NSFC) and the Shanghai Natural Science Foundation of China. His current research interests include IOT, GPU model for AI, and network computing. He is a Senior Member of the China Computer Federation (CCF), and a member of the Professional Committee of Open System of CCF and the Professional Committee of Computer Support Cooperative Work of Shanghai Computer Federation, China. He also served as a Program Committee Member of the 2008 IFIP International Conference on Network and Parallel Computing (NPC 2008).

LONGBING CAO is currently a Professor of information technology with the University of Technology Sydney (UTS), Australia. He is the Founding Director of the Advanced Analytics Institute, UTS. His primary research interests include data science and mining, machine learning, behavior informatics, agent mining, multi-agent systems, and open complex intelligent systems. He is currently dedicated to the research on non-iid learning in big data and behavior informatics, which involve very wide enterprise applications. He has successfully delivered 11 tutorials including IJCAI and CIKM, and dozens of invited talks to main conferences/workshops and public seminars to industry and government. He is the Chair of ACM SIGKDD, Australia, and the New Zealand Chapter, the IEEE Task Force on Data Science and Advanced Analytics, and the IEEE Task Force on Behavioral, Economic, and Socio-Cultural Computing. He serves as the Co-Chair for conference such as KDD2015, PAKDD13, and ADMA13, the Program Co-Chair or the Vice-Chair of PAKDD17, PAKDD11, and ICDM10, and the Area Chair or (senior) Program Committee Member of around 100 conferences including KDD, AAAI, IJCAI, ICDM, and AAMAS.

SONGLEI JIAN is currently pursuing the Ph.D. degree with the College of Computer, National University of Defense Technology, China, and the Advanced Analytics Institute, University of Technology Sydney, Australia. She has published over ten top-rank papers, including IJCAI, AAAI, and TKDE. Her research interests include representation learning, recommender systems, and unsupervised learning approach.

HAIYAN ZHAO received the Ph.D. degree from the Nanjing University of Science and Technology, in 2002. She is currently an Associate Professor with the Department of Computer Science and Engineering, University of Shanghai for Science and Technology. Her research interests include recommendation systems, workflow, and software testing.

JIAN CAO received the Ph.D. degree from the Nanjing University of Science and Technology, in 2000. He is currently a tenured Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU), China. He is the Leader of the Lab Service for Cloud Computing. He was a Postdoctoral Research Fellow of Shanghai Jiao Tong University (SJTU) and a Visiting Scholar with Stanford University. His research interests include intelligent data analytics, service computing, and network computing. He has published over 200 papers in referred conference and journals such as SIGKDD, IJCAI, AAAI, VLDB, WWW, INFOCOM, TMC, TOIS, TPDS, and TKDD.