




# CSTeller: forecasting scientific collaboration sustainability based on extreme gradient boosting

Wei Wang<sup>1</sup> · Bo Xu<sup>1</sup>  · Jiaying Liu<sup>1</sup> · Zixin Cui<sup>1</sup> · Shuo Yu<sup>1</sup> · Xiangjie Kong<sup>1</sup> · Feng Xia<sup>1</sup>

Received: 14 December 2017 / Revised: 10 February 2019 / Accepted: 2 June 2019 /  
Published online: 18 July 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2019

## Abstract

The mechanism why two strange scholars become collaborators has been extensively studied from the perspective of social network analysis. In academia, two scholars may collaborate with each other more than once, which means that scientific collaboration is to some extent sustainable. However, less research has been done to explore the sustainability of scientific collaboration. In this paper, we examine to what extent the collaboration sustainability can be predicted. For this purpose, an extreme gradient boosting-based collaboration sustainability prediction model named CSTeller is devised. We propose to analyze the sustainability of scientific collaboration from the perspectives of collaboration duration and collaboration times. We investigate factors that may affect collaboration sustainability based on scholars' local properties and network properties. These factors are adopted as input features of CSTeller. Extensive experiments on two real scholarly datasets demonstrate the effectiveness of our proposed model. To the best of our knowledge, this is the first attempt to explore scientific collaboration mechanism from the perspective of sustainability. Our work may shed light on scientific collaboration analysis and benefit many practical issues such as collaborator recommendation since a scientific collaboration is not a one-shot deal.

**Keywords** Scholarly big data · Deep learning · Relation mining · Coauthor network

## 1 Introduction

In modern academia, collaboration is often an important component of scientific research. Scientific collaboration brings scholars together to solve complex scientific problems [21]. Previous studies indicated that the scientific collaboration is becoming more and more

---

This article belongs to the Topical Collection: *Special Issue on Social Computing and Big Data Applications*

Guest Editors: Xiaoming Fu, Hong Huang, Gareth Tyson, Lu Zheng, and Gang Wang

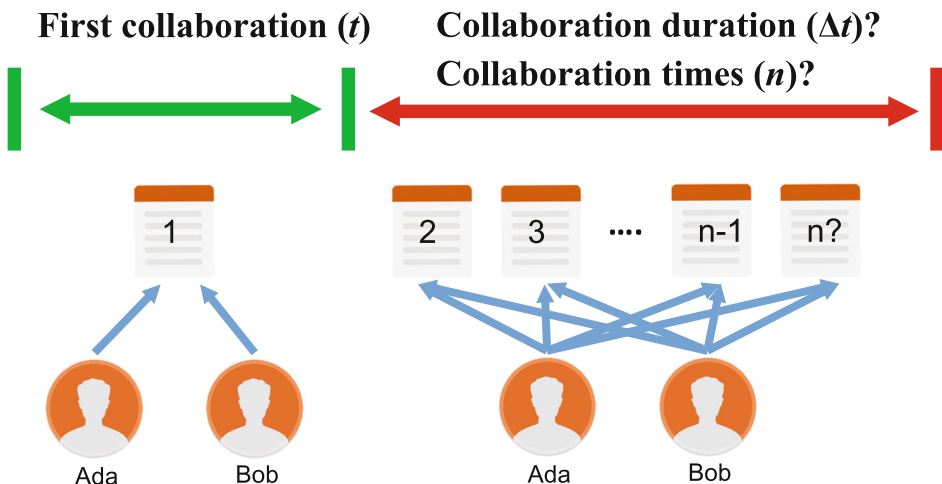
✉ Bo Xu  
boxu@dlut.edu.cn

<sup>1</sup> Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China

popular. Fruitful researchers tend to be more collaborative [42]. Due to the importance of scientific collaboration, many efforts have been done to understand the mechanism of scientific collaboration in order to promote scientific collaboration [14, 43, 52], i.e., collaborator recommendation [53].

Scientific collaboration mainly contains four stages including foundation, formulation, sustainment, and conclusion [42]. Most previous research focuses on analyzing factors that may affect the formulation stage [24, 45, 47, 49]. For example, the problem of link prediction has been extensively studied from the similarity perspective [24, 46]. However, what happens after a new collaboration is established? It is known to all that scientific collaboration is not a one-shot deal [16, 38]. It is not surprising that two scholars collaborate with each other more than once. In other words, scientific collaboration is sustainable (see Figure 1). The collaboration sustainability refers to the continuous state of scientific cooperation between two scholars. Such state is uncertain, which may be long term or short term. Specifically, we explore the collaboration sustainability from two perspectives, i.e., collaboration duration (CD) and collaboration times (CT), where CD stands for how long will a new collaboration last and CT stands for how many times will these two scholars collaborate with each other in the future (see Figure 1). Thus, how to find a sustainable collaborator? This work is different from the existing studies in link prediction [33, 46] and friend recommendation [28, 32]. Previous studies in scientific collaboration mainly focus on whether two scholars will collaborate with each other. We try to figure out the mechanism of sustainable collaboration and predict the sustainability of scientific collaboration.

Analyzing the sustainability of scientific collaboration is important. It has been proven that life partners resulted from sustainable collaboration have a significant impact on productivity and reputation [38]. It is not easy for scholars to find new collaborators. It is vital for scholars to maintain the academic network effectively. Scholars want to know once a new connection is built how long will it last and how often will it be. In this paper, we try to explore the following questions after the connection is established: 1) How long will this collaboration last? 2) How many times will this collaboration be? 3) Can the sustainability of this collaboration be predicted?



**Figure 1** An example of sustainable scientific collaboration

Predicting the sustainability of scientific collaboration is challenging. First, the nature of scholarly big data makes it difficult to extract needed factors [51]. The scale of scholarly data is nowadays very huge. Second, the CD and CT between scholars are uncertain. They follow the long-tail distribution (See Figure 2) [4]. Most collaborations will not last for a long time. It is difficult to build a prediction model with such unbalanced data [17, 31]. Third, factors affecting the sustainability of scientific collaboration are uncertain since few research has been done to analyze the mechanism of sustainable collaboration. Meanwhile, the interplay of many factors may confound the prediction performance.

To deal with the problems above, we propose a novel extreme gradient boosting model named CSTeller (Collaboration Sustainability Teller) to predict the sustainability of a new collaboration based on the local and network properties [9]. In the scenario of collaboration sustainability prediction, it is important to consider not only the topology of scientific collaboration network, but also scholars' various academic characteristics. While previous works mainly adopt network metrics, CSTeller proposes to profile scholars from two groups of factors, including personal factors (i.e., academic age, degree, and publication counts) and social factors (i.e., common neighbors and shortest path). Due to large-scale nature of scholarly datasets, the CSTeller is designed based on the framework of Extreme gradient boosting, which has been proven effective its its scalability in many scenarios [1, 2]. We evaluate the performance of CSTeller based on two scholarly datasets extracted from DBLP where experimental results show that CSTeller outperforms benchmark machine learning methods. Our major contributions can be summarized as:

- **Problem Formulation.** We formulate the problem of collaboration sustainability prediction from the perspectives of the CD prediction and the CT prediction.
- **Feature Selection.** We crawl scholars' local properties including academic ages, number of publications, and number of collaborators, as well as network properties including degree and shortest path.

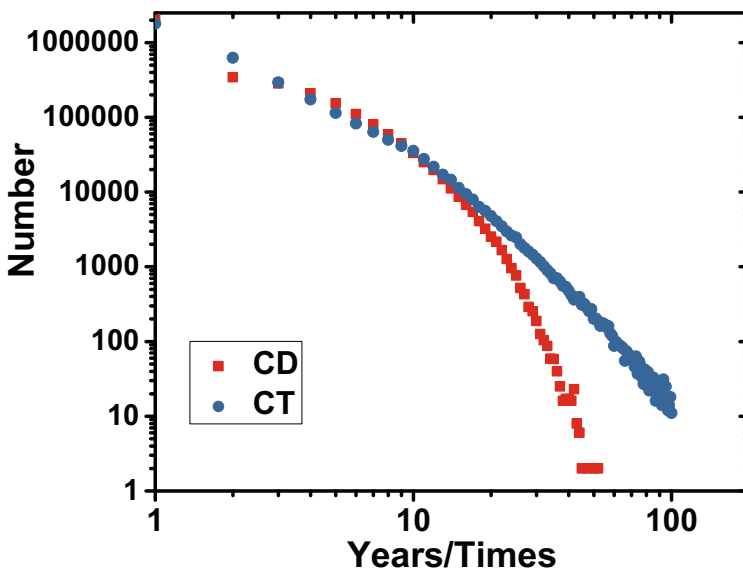


Figure 2 Distributions of scientific collaboration duration and times

- **Prediction Algorithms.** We present a tree-boosting-based collaboration sustainability prediction model called CSTeller, and conduct extensive experiments on two datasets to verify the performance of the proposed model.
- **New Insight.** We propose to study the mechanism of scientific collaboration from the perspective of sustainability, which will shed light on collaborator recommendation.

The rest of this paper is organized as follows. Section 2 reviews the related work. We formulate the sustainability prediction problem in Section 3. The factors that may determine the sustainability of scientific collaboration are discussed in Section 4. We present our proposed model in Section 5. Experimental results are shown in Section 6. Finally, Section 7 concludes this paper.

## 2 Related work

Predicting the collaboration sustainability relies on the power of scholarly big data analysis. In recent years, scholars are producing an increasing number of publications [7, 22]. It leads to the emerging of a new research area, scholarly big data [23, 54].

Scientific collaboration, as a research topic, has been investigated by many scholars in diverse disciplines including information science, social science, computer science, as well as any discipline where scientific collaboration happens [34, 56]. Scientific collaboration is becoming more and more popular because it has the potential to solve complex scientific problems and promote scientific research [42]. Previous studies have shown the continuous increase in the number of co-authored articles in many disciplines within and across institutions and countries [11]. Meanwhile, co-authored papers will gain more citations than single-authored papers [37]. A single scholar may not possess all the expertise or information needed to tackle a complex scientific issue. Funding agencies are promoting interdisciplinary, inter-institution, and international collaboration.

Scientific collaboration network extracted from coauthor relationship is a typical way to reveal the collaboration patterns [49, 50]. For example, social scientists have used quantitative methods to investigate the mechanism of scientific collaboration based on co-authored networks [35, 38]. Newman [35] presented the first investigation on the collaboration network by analyzing different network properties of scientific collaboration network such as clustering, giant component, centrality, and shortest path. From the career path perspective of scientific collaboration network, Peterson analyzed 166,000 collaboration records and found that scientific collaboration networks are dominated with weak collaboration relationship characterized by high turnover rates [38]. However, although tremendous efforts have been done to analyze the scientific collaboration mechanism, few works have been done to investigate the sustainability of scientific collaboration.

There are various algorithms that can be used to do predictions. Many machine learning tools have been developed. Support Vector Machines [55], Decision Trees [39], Linear Regression [40], K-Nearest Neighbors [27], and Random Forests [5] are some of the popular algorithms used for prediction. Although these classical methods have achieved great success in prediction, each of them has its own shortcomings and practical constraints in terms of prediction accuracy and time consuming. Our CSTeller model is inspired by the Xgboost model [9], which is an efficient and scalable variant of Gradient Boosting Machine. It has been proven to be a powerful tool for several data mining competitions [1].

Few works have been done to reveal the collaboration mechanism after a collaboration has been established [6, 8]. Scientific collaboration has been studied from the perspective

of network science [18, 36]. Many efforts have been done to predict the collaborative relationships from the perspective of link prediction [3, 30, 44]. Meanwhile, the dynamics of scientific collaboration has been studied from the perspective of time-aware link prediction based on evolving networks [10, 19, 20, 25].

In reality, collaboration is not a one-shot deal where two scholars may collaborate with each other more than once. Since it is not easy to find a suitable collaborator, we would like to know the sustainability of a connection. In this paper, we try to explore the following questions after the connection is established: 1) How long will this collaboration last? 2) How many times will this collaboration be? 3) Can the sustainability of this collaboration be predicted? To tackle these issues, we proposed CSTeller, which is a tree-boosting based supervised machine learning method, to predict the sustainability of scientific collaboration. Since few work has been done to explore the mechanism of sustainable collaboration, our work will shed light on collaboration analysis and collaborator recommendation. This work is extended from our previous poster paper [50], which is a preliminary work on collaboration sustainability prediction which focuses on feature engineering and regression method selection.

### 3 Problem definition

Typically, the task of scientific collaboration sustainability prediction can be formulated as a regression problem for predicting CD and CT. However, the long-tailed distributions of CD and CT (see Figure 2) make such prediction inevitably challenging. Meanwhile, the collaboration between two scholars is not static. The CD may last many years as  $\Delta t$  and they may collaborate  $m$  times during the CD  $\Delta t$ . From that perspective, we need to infer the collaboration records after the first collaboration between a scholar pair. It is worth mentioning that the collaboration sustainability may not be term “long-term collaboration relationship” because two scholars may merely collaborate few times (i.e., two times) and such collaboration may last for few years (i.e., one year). In other words, the collaboration sustainability between two scholars is uncertain.

We formulate the collaboration sustainability prediction issue as two prediction problems, namely, CD prediction and CT prediction. We assume a collaboration pair  $i$  and  $j$ . We define and calculate a set of factors  $\{x_1, x_2, x_3, \dots, x_n\}$  that determine the collaboration sustainability. For example,  $x_1$  can be the shortest path in the entire scientific collaboration network between these two scholars. The task becomes finding two suitable models  $f(x, y)$  to describe the  $y_1$  (CD) and  $y_2$  (CT) separately, where the  $y_1$  and  $y_2$  denote the dependent variables of factors  $x$ .

With the above analysis, our problems can be formally defined as follows:

**Given:** The collaboration records between a scholar pair  $i$  and  $j$  extracted from the DBLP digital library when these two scholars begin their collaboration.

**Predict:** The collaboration sustainability of this collaboration in terms of CD and CT.

CD prediction and CT prediction are two different problems because the distribution of these two phenomena is different. For example, the collaboration may last at most 45 years and the CT may be more than 100 (see Figure 2).

In order to solve these two proposed problems, we firstly analyze the factors that are closely related with collaboration sustainability in the following section. Then, based on these factors, we design our prediction model.

**Table 1** Factor definitions

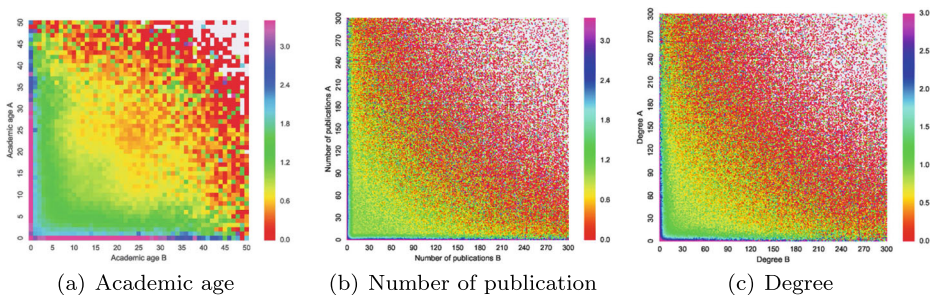
Factors	Description
Academic Age (AA)	Academic ages of A and B when first collaborating
Number of Publications (NP)	Number of publications of A and B before collaboration
Degree (DG)	Number of collaborators of A and B before collaboration
Common Neighbors (CN)	Number of common neighbors of A and B before collaboration
Shortest Path (SP)	Shortest path between A and B before collaboration

## 4 Empirical analysis of collaboration sustainability

We perform an empirical analysis of the factors that may influence the collaboration sustainability extracted from the DBLP dataset. There are various factors that may drive the collaboration sustainability such as geographical position, collaboration preference, and research interest. It is known that more factors may bring better prediction results. A few work has been done on collaboration sustainability prediction [20, 48]. However, they mainly focus on link prediction whereas the nature of scientific collaboration is overlooked. Due to the data limitation and for simplicity, in this paper, we mainly explore two critical groups of factors including personal factors and social factors, as shown in Table 1. The first three factors Academic Age (AA) [50], Degree (DG), and Number of publication (NP) are personal factors. The last two factors Common Neighbors (CN) and Shortest Path (SP) are the social factors. Note that all these factors are calculated or extracted based on the scientific information network before the collaboration has been established. The dataset used in this section is the largest giant component of the scientific collaboration network in which each scholar has more than 10 publications.

### 4.1 Personal factors

The prediction task for collaboration sustainability between two scholars naturally depends on the scholars themselves. Personal factors play an important role in both establishing and maintaining scientific collaboration. Personal factors including academic reputation,



**Figure 3** The impact of scholar's personal factors on CD. The x and y axes represent a academic age, number of publications, and degree between any two collaborative scholars, respectively

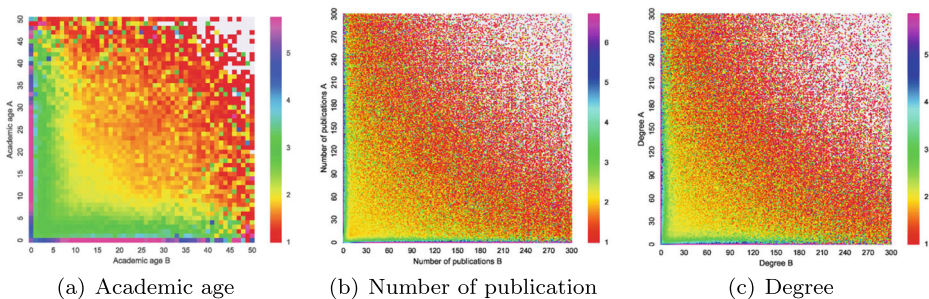
collaboration preference, and career stages may influence scholars' collaboration behaviors greatly.

Figure 3 depicts the relationship between scholars' personal factors and collaboration sustainability in terms of CD. In Figure 3a the color of each pixel represents the average value of CD between two scholars in different AA. The maximum AA considered in this paper is 50. The graph shows that there has been a marked decrease in the CD with the increase of AA. The collaboration with beginning scholars ( $AA \leq 5$ ) will last longer while the collaboration between two senior scholars may last only one year. The impact of NP on CD is shown in Figure 3b, where the maximum NP considered in this paper is 300. We can see that CD sharply decreases with the increase of NP. Similar trends can be seen from Figure 3c, where the maximum DG is 300. From Figure 3a, b, and c, we can find out that the collaboration with beginning scholars who has small AA, NP, and DG will last longer. With the increase of these personal factors, the CD is likely to decline. The impact of personal factors on CT can be seen from Figure 4a, b, and c. The overall trends of CT are similar to CD, where the CT with beginning scholars will last longer and CT declines obviously with the increase of AA, NP, and NG.

These findings are consistent with the practical situation, which is in line with the findings in weak tie phenomenon where two scholars are in a weak relationship if they are not familiar with each other [15]. For example, the collaborations between an advisee who has a smaller AA (NP and DG) and his/her advisor who has a bigger AA (NP and DG) is in a stable condition and will last for a long time. Usually, an advisee will collaborate many times with his/her advisor in the process of pursuing a Ph.D. degree. On the contrary, the collaboration between two senior scholars will last a short time according to Figures 3 and 4.

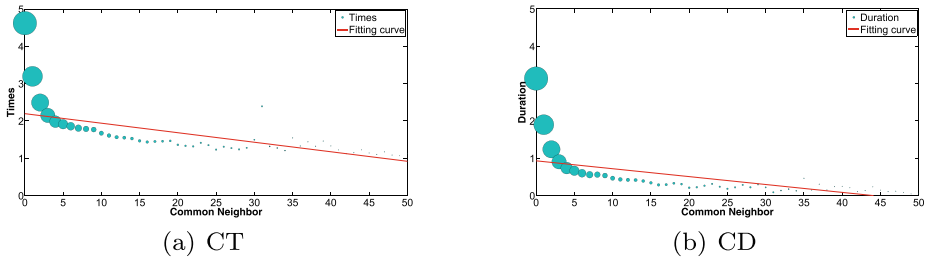
## 4.2 Social factors

Aside from the personal factors of scholars, another intuitive factor affecting a collaboration's sustainability is the social relationship between a collaboration pair. Previous studies have shown that the social position of a scholar has great impact on his/her academic performance [12]. We assume that the collaboration sustainability will be influenced by the social factors. To explore this assumption, we construct a large scientific collaboration network from DBLP dataset, where each node represents a scholar and two nodes are considered connected if the scholars have collaborated with each other. We then extract two simple



**Figure 4** The impact of scholar's personal factors on CT. The x and y axes represent a academic age, number of publications, and degree between any two collaborative scholars, respectively





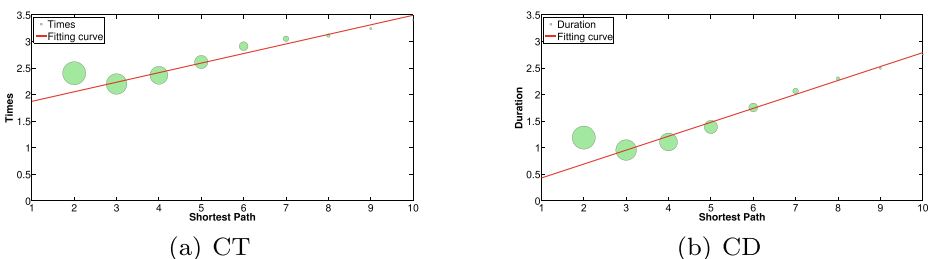
**Figure 5** Impact of number of common neighbor between scholars on collaboration sustainability

and basic features between two collaborators from the collaboration network including the shortest path and common neighbors.

Figure 5 presents the results of social factor analysis. The size of each circle represents the number of scholars, where a bigger size means that there are more scholars in this condition. The red line in each subfigure is the fitting curve. The maximum CN and SP considered in this paper are 50 and 10 respectively. Figure 5 describes the impact of CN on collaboration sustainability. One unanticipated finding is that with the increase of CN, both the CD and CT decrease accordingly. Meanwhile, as it can be seen from Figure 6, with the increase of SP, the CD and CT increase apparently. In reality, most collaborator recommendation method takes advantages of CN or random walk to find suitable collaborator candidates. The famous social theory triadic closure [26] also indicates that people with common neighbors tend to become friends. However, our findings suggest that the collaboration established from less social similarity will be more sustainable. In other words, the collaboration between scholars with close social relationship will not last long.

## 5 Design of CSTeller

In this section, we describe our predictive model CSTeller to forecast the scientific collaboration sustainability when two scholars collaborate with each other for the first time. Specifically, we make predictions both on CD and CT. We first give an overview of CSTeller. Then, we present each section in detail including tree ensemble, gradient boosting, and feature extraction. Moreover, we give an example showing how to calculate the input features from the DBLP dataset.



**Figure 6** Impact of shortest path between scholars on collaboration sustainability



## 5.1 Overview of CSTeller

The CSTeller sustainability prediction model is inspired by the fact that the child is the father of the man [29], which means that the collaboration sustainability can be forecasted at the early stage. In this paper, we define the early stage as the time point when two scholars begin collaborating with each other. Meanwhile, since extreme gradient boosting has been proven to be a powerful machine learning algorithm both for classification and regression [9], we adopt it as our foundation. Furthermore, we design many scholarly features for model training. The overall framework of CSTeller is shown in Figure 7.

- All input features are extracted from the DBLP digital library. The original data is a set of papers published by scholars in the field of computer science. In order to eliminate authors who do research only for a short time we limit our research to scholars who have published at least 10 papers [41]. We reconstruct the collaboration profile of scholars and gain the collaboration records of any two co-authored scholars. The personal factors are extracted from the meta data including academic ages, number of publications, and number of coauthors.
- To extract the social factors, we need to construct the collaboration network, where two scholars are regarded connected if they have co-authored at least one paper. Meanwhile, in order to filter out those isolated nodes, we extract the largest connected component of the collaboration network. Based on this largest connected component, we can calculate the social factors.
- Since the sustainability prediction is a regression task. The CSTeller conducts a series of decision trees which is trained with the gradient boosting approach. The tree ensemble and gradient boosting sections are detailed introduced in the following section.
- The sustainability of scientific collaboration is studied from two perspectives including CD and CT. Thus, the CSTeller model will do prediction on these two issues.
- Finally, we evaluate the performance of CSTeller with four typical evaluation metrics.

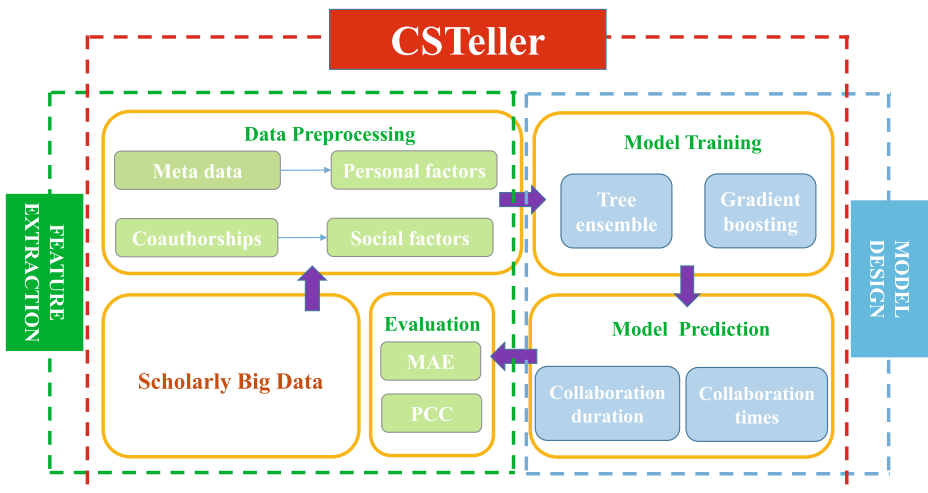


Figure 7 Framework of CSTeller

**Algorithm 1** Pseudocode of addictive training of CSTeller.

**Input:** Training set  $\{(x_i, y_i)\}_{i=1}^n$ ;  
 Loss function  $L(y, F(x))$ ;  
 Number of interaction  $M$

**Output:**  $F_M(x)$ ;

1: Initialize model with a constant value:

$$F_0(X) = \arg \min \sum_{i=1}^n L(y_i, \gamma).$$

2: For  $m = 1$  to  $M$ :

1. For  $i = 1, \dots, n$ , compute the pseudo-residuals:

$$\gamma_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}x}$$

2. Fit a base learner  $h_m(x)$  to pseudo-residuals;

3. Train it using the training set  $\{(x_i, r_m)\}_{i=1}^n$ ;

4. Compute multiplier  $\gamma_m$  by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg \min \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)).$$

3: Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x).$$

## 5.2 Tree ensemble

The CSTeller model aims to make the prediction of  $y_i$  ( $y_i$  can be the CD or times). Given  $x_i$ , where  $x_i$  represents the input features, the prediction task is to find the best parameters given the training data. In order to find the best parameters that can better describe the data, people always define a so-called objective function, which usually contains two parts, training loss and regularization:

$$Obj(\Theta) = L(\theta) + \Omega(\theta) \quad (1)$$

where  $L$  is the training loss function and  $\Omega$  is the regularization term. The training loss function  $L$  measures the performance of proposed model on training data and the regularization term  $\Omega$  controls the complexity of the model which helps to avoid overfitting.

Similar with XGboost, CSTeller is an ensemble of a set of classification and regression trees (CART) [2]. The prediction scores of each CART are summed up to get the final score, which can be calculated as follows:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (2)$$

where  $K$  is the number of ensemble trees,  $f_k$  means an independent tree, and  $F$  is the set of all possible CARTs. Therefore, we can rewrite (1) as follows:

$$Obj(\Theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3)$$

The regularization term  $\Omega$  in CSTeller is given by:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (4)$$

where  $T$  and  $\omega$  are the number of leaves and their corresponding scores.  $\gamma$  and  $\lambda$  are parameters controlling the degree of regularization.

### 5.3 Gradient boosting

Since (4) uses a function as parameter, it cannot be optimized with traditional optimization methods in equation space. Thus, we train the model in an additive manner. Let  $\hat{y}_i^{(t)}$  be the predictive result of the  $i$ -th instance at the  $i$ -th iteration, we have:

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ &\dots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned} \tag{5}$$

The additive training method is shown in the algorithm 1. We add  $f_t$  to optimize the following objective function:

$$\Upsilon^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{6}$$

We take Taylor expansion [9] of the objective and define  $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$  and  $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ . Thus we can rewrite (6) as follows:

$$\begin{aligned} \hat{\Upsilon}^{(t)} &= \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \\ &= \sum_{j=1}^T [(\sum_{i \in I_j} g_i) \omega_j + \frac{1}{2} (\sum_{i \in I_j} h_i) \omega_j^2] + \gamma T \end{aligned} \tag{7}$$

where  $I_j = \{i | q(x_i) = j\}$  represents the instance set of leaf  $j$ . Thus the optimal leaf weight  $\omega_j^*$  can be calculated as:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{8}$$

where  $G_j = \sum_{i \in I_j} g_i$  and  $H_j = \sum_{i \in I_j} h_i$ . The resulting objective value can be calculated as:

$$Obj = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \lambda T \tag{9}$$

In this condition, a smaller  $Obj$  means a better tree structure.

### 5.4 Learn the tree structure

Meanwhile, for each leaf node of the tree, we need to add a split. The change of the objective after splitting can be calculated as:

$$Gain = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{10}$$

where  $\frac{G_L^2}{H_L+\lambda}$  stands for the score on the left leaf,  $\frac{G_R^2}{H_R+\lambda}$  stands for the score on the right leaf,  $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$  stands for the score on the original leaf without splitting, and  $\gamma$  represents the regularization on the additional leaf.

## 5.5 Feature extraction

We used all these personal and social factors (as shown in Table 1) as the input features to predict the collaboration sustainability:

### Personal factors:

- AA: AA means the academic ages of scholars A and B when they firstly collaborated with each other. We take advantage of the factor of AA based on the fact that scholars tend to have different collaboration strategies at different career stages [38]. Apparently, a PhD candidate will collaborate frequently with his/her advisor.
- NP: NP means the number of publications of scholars A and B when they collaborate for the first time with each other. A scholar's publications can, to some extent, reflect his/her academic performance. Fruitful scholars tend to be more collaborative and may have a higher reputation.
- DG: DG means the number of collaborators of scholars A and B when they firstly collaborate with each other. Similar with NP, DG can also reflect the collaboration strategies for different scholars.

### Social factors:

- CN: CN means the number of common neighbors of scholars A and B before they collaborate with each other. Based on the famous social theory triadic closure [26], people share more common neighbors tend to be connected in the future. Thus, we adopt the CN to measure how similar two scholars are in the network.
- SP: SP means the shortest path between scholar A and B in the scientific collaboration network before they collaborate with each other. The SP is used to measure how close two scholars are in the network.

The academic age is calculated by the investigated year minus the year he/she firstly publishes a paper. All the input features are normalized into [0, 1] in order to improve the learning efficiency. The normalized method we adopt in this paper is the min-max normalization:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (11)$$

Note that all the input features are calculated exactly the time when two scholars begin their collaboration. In order to calculate the shortest path between two scholars, we construct the collaboration network for each collaboration. For example, if scholar A and B begin their collaboration in the year 2000, we will extract all papers published before that. It cannot be as accurate as day or month because the exact time point of the month or day is incomplete in the row data. Then, the collaboration network  $Network_{AB}$  is constructed based on the coauthorship extracted from these papers. Finally, the shortest path is calculated based on the largest connected component of this collaboration network  $Network_{AB}$ .

Figure 8 shows a brief example of how we extract all the input features from the collaboration records in DBLP. In this figure, we aim to predict the collaboration sustainability between scholar Linda and scholar Bob. Linda begins her research from 2014 and she has four papers co-authored with three collaborators Feng, Wei, and Ivan. Bob begins his research from 2005 and he has three papers co-authored with two collaborators Feng and Ivan. Meanwhile, Linda and Bob begin their collaboration in 2015. Thus, the AA of Linda and Bob are 2 and 10 respectively. The NP of Linda and Bob are 4 and 3 respectively. The DG of Linda and Bob are 4 and 2 respectively. On the other hand, from their collaboration records, we can construct the collaboration network in 2015. Note that many other scholars are considered to construct the collaboration network to calculate the social factors between Linda and Bob. For simplicity, we merely show the basic collaboration network in this figure. Thus, the CN between Linda and Bob are 2 including Feng and Ivan. The SP between Linda and Bob is 2. Finally, we can gain all the input features to build our prediction model.

## 6 Performance evaluation

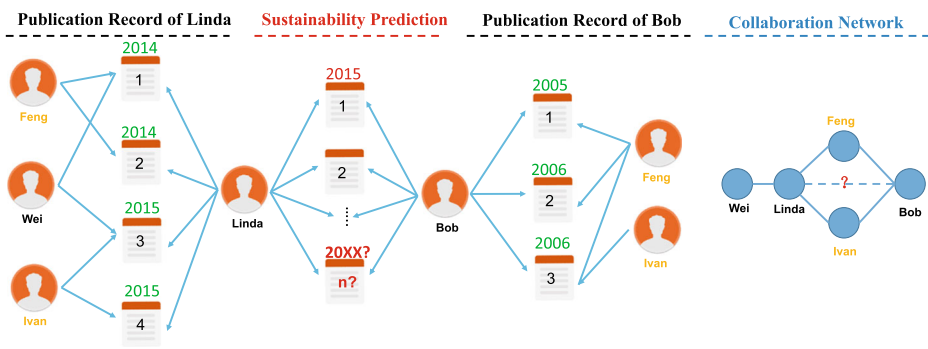
In this section, we design extensive experiments to evaluate the performance of CSTeller with two real datasets. Since this is the first work proposed to predict the collaboration sustainability, there is not too much room for comparison algorithms. Therefore, we compare our model with the typical machine learning method with four popular evaluation metrics.

Meanwhile, in order to investigate the contribution of each input feature on the performance of CSTeller, we employ the “jackknife” [13] method with three cases: (1) Removing one factor and predicting with the rest factors (Removing); (2) Using only one factor to do prediction (Adding), and (3) Predicting with all factors (All).

All experiments are performed on a 64-bit Windows-based operation system, with a 4-duo and 2.6-GHz Intel Xeon CPU, 128-G Bytes memory. All the experiments are realized with Python.

### 6.1 Datasets

We extract two distinguished investigated groups from the DBLP and APS datasets, respectively. These two scholarly datasets record the meta data of a paper. The DBLP dataset



**Figure 8** An example of feature extraction

indexes more than 2.3 million articles in the field of computer science. The dataset contain the meta data of papers including authors, title, pages, years, crossref,(in) proceedings or journals, etc. The DBLP dataset can be freely accessed from <http://dblp.dagstuhl.de/xml/>. The APS dataset is comprised of over 450,000 articles dating back to 1893 in the field of Physics. The APS dataset includes DOI, journal, volume, issue, first page and last page or article id and number of pages, title, authors, affiliations, publication history, PACS codes, table of contents heading, article type, copyright information, and citation relationships. Researchers can request access to the APS dataset by filling out a simple Web form from <https://journals.aps.org/datasets>. Based on the meta data of these two datasets, we can construct the scientific collaboration network and calculate all the input features.

Since the APS dataset does not provide unique author identifiers, two distinct scholars may have the same full or short name. To this end, we conduct a comprehensive name disambiguation process based on the idea in [41]. Meanwhile, to exclude authors who leave academia at their early academic career, we limit our analysis to scholars who (1) have published at least 10 papers, (2) have published at least one paper every 5 years, (3) their first collaboration should happen at least 20 year before 2016, (4) have no publication record by 2011 (five years before 2016). The first three principles are designed to limit our research on authors who are active in academia. The last two principles are proposed to ensure a sufficient time to calculate the collaboration sustainability between two scholars after first collaboration. The statistics of these two datasets can be seen from Table 2. We can see that the collaboration among scholars in APS are more sustainable.

## 6.2 Evaluation metrics

Collaboration sustainability prediction is a regression problem instead of classification. In a regression problem, we need to predict a series of continuous value. Thus, in order to evaluate the performance of CSTeller, we adopt four typical metrics including MAE (Mean Absolute Error), MSE (Mean Square Error), PCC (Pearson's Correlation Coefficient), and CCC (Concordance Correlation Coefficient). Given the true value of  $y$  ( $y$  can be CD or times), and the predictive value  $\hat{y}$ , the MAE is given by:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (12)$$

The MSE is given by:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (13)$$

The PCC is given by:

$$PCC = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \quad (14)$$

**Table 2** Statistics of two investigated groups

Group	Nodes	Edges	Duration	Times
DBLP	185739	3443845	2.669	2.983
APS	14022	355992	3.077	3.825

where the  $\bar{y}$  is the mean of  $y$  and  $\bar{\hat{y}}$  is the mean of  $\hat{y}$ .

The CCC is given by:

$$CCC = \frac{2s_{y\hat{y}}}{s_y^2 + s_{\hat{y}}^2 + (\bar{y} - \bar{\hat{y}})^2} \tag{15}$$

where  $s_{y\hat{y}}$  is the covariance between  $y$  and  $\hat{y}$ ,  $s_y^2$  and  $s_{\hat{y}}^2$  are the variances of  $y$  and  $\hat{y}$  respectively. From the definitions of these metrics, we can see that a better prediction results will have low MAE and MSE, and high PCC and CCC.

### 6.3 Baseline method

To our knowledge, the specific issue we address has not been tackled before. Hence, we selected strong machine learning algorithms for regression. Specifically, we compare our proposed CSTeller with a series standard regression model, including Linear Regression (LR) and Support Vector Machine (SVM). Meanwhile, we compare CSTeller with Time-aware Link Prediction (TLP) [10] which is a state-of-the art method for evolving link prediction. Specifically, TLP is a supervised classification methods considering network topology similarity metrics.

In the experiments, we use all the potential factors as the input feature in all the comparison method. Generally, we illustrate the prediction result of each method on two datasets to show the predictability of collaboration sustainability, whereas we only use CSTeller to explore the factor contribution with “jackknife” approach.

### 6.4 Effect of training data size

We perform our experiments on two different research groups. We divide each group into two subsets, the training set and the testing set, where the training set is used to train the parameters of our model and the testing set is used to evaluate the performance of the proposed model. Specifically, we randomly select twenty percentage of the data of each group as the testing set. In order to explore the performance of the CSTeller in terms of training data size, we perform experiments on different fractions of training data size ranging from

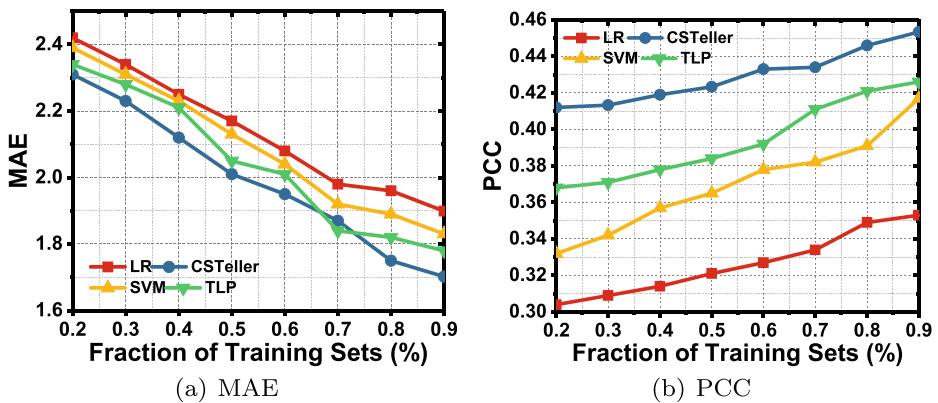


Figure 9 Performance of CSTeller and baseline methods on CD prediction over different training sets with DBLP dataset



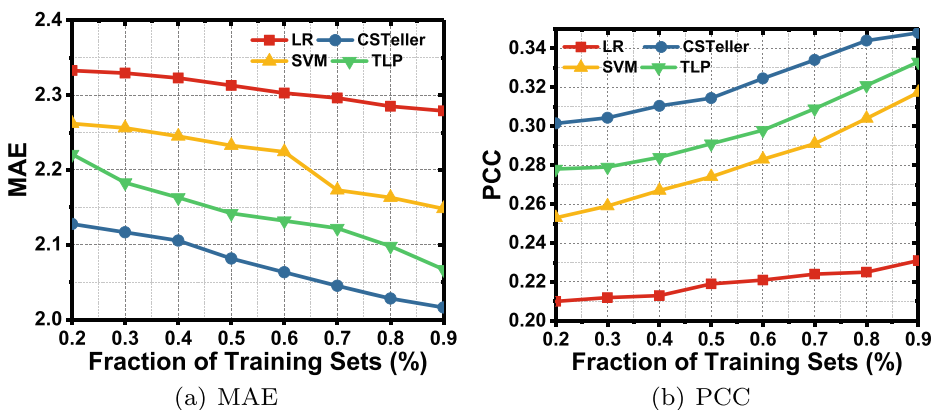
10% to 90%. Meanwhile, the k-fold cross validation is adopted in all experiments in order to enhance the stability and fidelity of our model.

Figure 9 depicts the performance of CSTeller and baseline methods on CD prediction in terms of different fractions of training set on DBLP dataset. From the Figure 9a, we can see that the MAE of all methods declines with the increasing of fractions of training set, which means that prediction task will benefit from large amount of training data. When the fraction of training set goes up from 20% to 90%, take DBLP for example, the MAE of CSTeller decreases from 2.32 to 1.69 and the MAE of LR decreases from 2.45 to 1.91. On the other hand, we can observe from this figure that the proposed model CSTeller always has better performance than baseline methods. In particular, CSTeller outperforms 11 and 12 percentage on DBLP and APS datasets respectively in terms of MAE compared with LR method.

Moreover, CSTeller always achieves better results on CD prediction than LR in terms of PCC (Figure 9b). From this figure, we can get the conclusion that with the increasing fractions of training data, all prediction methods will achieve better CD prediction results and CSTeller always outperforms baseline methods in terms of MAE and PCC. As discussed before, the prediction of collaboration sustainability contains not only the CD prediction but also the CT prediction. Figure 10 illustrates how the fractions of training set influence the performance of CSTeller and baseline methods with CT prediction on DBLP dataset. Figure 10a shows the CT MAE of CSTeller and baseline methods on DBLP and dataset. We can see that with the increasing fractions of training set, all method will achieve better performance, which is similar to the trend of CD prediction. Meanwhile, the MAE of CSTeller is always lower than other methods, which means that CSTeller has better prediction results. Another observation is that with the increasing fractions of training set, the MAE of CSTeller decreases faster than other methods, which shows that CSTeller can better take advantages of a larger dataset.

## 6.5 Results on different datasets

In order to evaluate the performance of CSTeller in terms of different datasets, we perform experiments on two different research groups (see Table 2). Specifically, the scholars



**Figure 10** Performance of CSTeller and baseline methods on CT prediction over different training sets with DBLP dataset

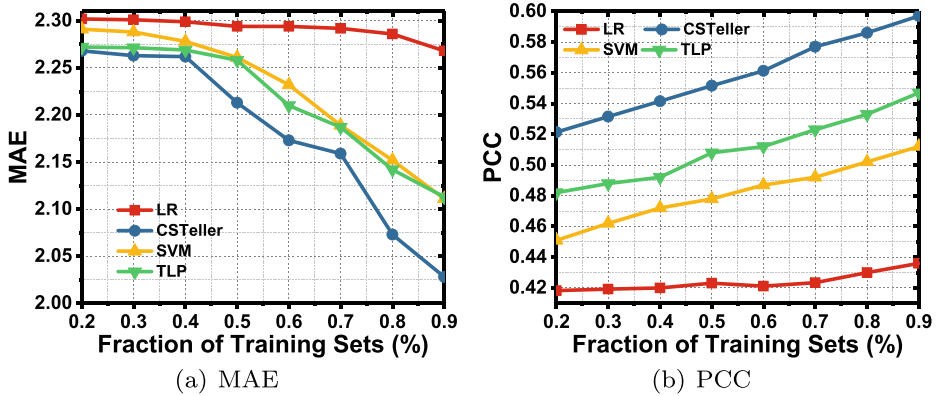


Figure 11 Performance of CSTeller and baseline methods on CD prediction over different training sets with APS dataset

in DBLP have at least 10 publications indexed in the DBLP digital library and the scholars in APS have at least 80 publications indexed. From Table 2, we can see that both CD and CT among scholars in APS are higher than scholars in DBLP. Thus, we can evaluate the performance of CSTeller more comprehensively by running experiments on these two distinguished datasets. The results on APS dataset are illustrated in Figures 11 and 12. We can see on these two figures that similar with the results on DBLP dataset, CSTeller always achieve the best performance than baseline methods. By comparing the results on DBLP and APS datasets, we can easily find that CSTeller can better predict the CD on APS than on DBLP in terms of MAE and PCC. The reason is that the collaboration relationships in APS dataset is more stable.

### 6.6 Factor contribution analysis

In order to predict the collaboration sustainability among scholars, we have introduced two groups of factors including personal factors and social factors.

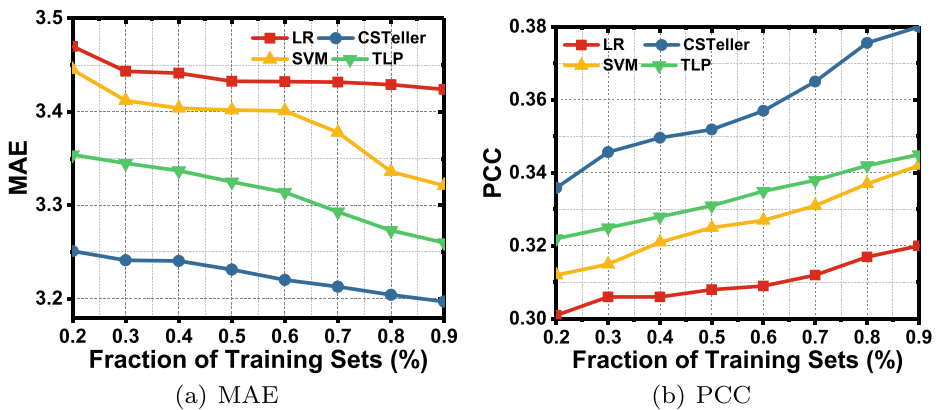


Figure 12 Performance of CSTeller and baseline methods on CT prediction over different training sets with APS dataset

**Table 3** Factor contribution analysis on CD prediction with DBLP dataset

CD of DBLP	Rules	MAE	MSE	PCC	CCC
AA	Removing	1.721	7.568	0.431	0.312
	Adding	1.905	8.302	0.322	0.186
NP	Removing	1.724	7.564	0.431	0.311
	Adding	1.891	8.285	0.334	0.201
CN	Removing	1.867	8.097	0.365	0.229
	Adding	1.826	8.240	0.335	0.201
SP	Removing	1.727	7.591	0.428	0.309
	Adding	1.932	8.425	0.347	0.174
DG	Removing	1.718	7.544	0.338	0.309
	Adding	1.884	8.157	0.275	0.216
	All	1.702	7.044	0.453	0.332

To explore the contributions of these factors, we adopt the “jackknife” approach [13] with three cases: (1) Removing one factor and predicting with the rest factors (Removing); (2) Using only one factor to do prediction (Adding), and (3) Predicting with all factors (All). Based on these strategies, we can find out the individual contribution that each factor supports to the overall prediction task. Tables 3, 4, 5, and 6 show the MAE, MSE, PCC, and CCC for the three cases with different research groups including CD and CT prediction. We can see that the contribution of each factor has different influence.

Table 3 shows the CD prediction results with “jackknife” approach on DBLP dataset. In Table 3, 10 percentage drops (from 1.702 to 1.867) in MAE value by the removal of CN factor (Removing strategy) which shows that the CN factor plays an important role in predicting the CD. At the same time, the relative little decrease by the removal of other input factors indicates that remaining factors provide limited contributions in CD prediction. When we take advantages of the Adding strategy, the CN factor still achieves the best performance, though the DG factor also has a remarkable effect on the CD prediction. Similar results can be seen on MSE, PCC, and CCC. Furthermore, when using all the factors, the

**Table 4** Factor contribution analysis on CT prediction with DBLP dataset

CT of DBLP	Rules	MAE	MSE	PCC	CCC
AA	Removing	2.123	19.405	0.341	0.205
	Adding	2.255	19.770	0.271	0.137
NP	Removing	2.130	19.516	0.335	0.197
	Adding	2.221	20.415	0.283	0.148
CN	Removing	2.211	19.718	0.301	0.161
	Adding	2.258	20.472	0.242	0.110
SP	Removing	2.123	19.236	0.342	0.206
	Adding	2.303	21.115	0.236	0.105
DG	Removing	2.124	19.642	0.338	0.202
	Adding	2.238	20.021	0.275	0.141
	All	2.016	18.943	0.348	0.209

**Table 5** Factor contribution analysis on CD prediction with APS dataset

CD of APS	Rules	MAE	MSE	PCC	CCC
AA	Removing	2.066	10.799	0.520	0.418
	Adding	2.272	12.134	0.423	0.300
NP	Removing	2.068	10.844	0.522	0.420
	Adding	2.216	11.455	0.452	0.343
CN	Removing	2.179	11.050	0.480	0.371
	Adding	2.237	12.503	0.362	0.232
SP	Removing	2.090	10.964	0.509	0.407
	Adding	2.385	12.645	0.353	0.229
DG	Removing	2.068	10.847	0.508	0.408
	Adding	2.207	11.380	0.473	0.362
	All	2.028	10.617	0.597	0.421

CSTeller will achieve the best performance, which indicates that all the considered factors are useful in predicting the CD with DBLP dataset.

The CT prediction results with “jackknife” approach on DBLP dataset are shown in Table 4. From this figure, we can see that the CN factor still plays the most important role in predicting the CT. Specifically, when using the Removing strategy with the CN factor, the MAE is 2.11, which is the highest MAE among other Removing strategies. That is to say, the CN factor is most closely related to the CT prediction. When using the Adding strategy, the factor NP can achieve the best result on MAE, which means that the NP factor can be used to better predict the CT alone. Meanwhile, the All strategy always achieves the best performance in terms of MAE, MSE, PCC, and CCC, which indicates that all the considered factors are useful in predicting the CT with DBLP dataset.

Table 5 shows the results of CD prediction on APS dataset. From this table, we can see that all the selected factors are important in predicting the collaboration for scholars who have fruitful publications both by Removing strategy and Adding strategy. Different from the CD prediction for DBLP in Table 3, the factor DG plays the most important role

**Table 6** Factor contribution analysis on CD prediction with APS dataset

CT of APS	Rules	MAE	MSE	PCC	CCC
AA	Removing	3.206	59.824	0.357	0.234
	Adding	3.357	62.620	0.318	0.189
NP	Removing	3.216	61.951	0.360	0.232
	Adding	3.306	59.513	0.336	0.212
CN	Removing	3.324	66.151	0.341	0.206
	Adding	3.348	62.066	0.229	0.103
SP	Removing	3.212	62.046	0.361	0.232
	Adding	3.446	66.657	0.265	0.139
DG	Removing	3.211	65.295	0.357	0.226
	Adding	3.332	63.739	0.331	0.197
	All	3.199	58.717	0.370	0.240

observed from the Adding strategy. Meanwhile, the All strategy still has the best performance in terms of MAE, MSE, PCC, and CCC compared with the Adding and Removing strategies.

The results of CT prediction on APS dataset are shown in Table 6. Similar to the results in Table 4, the CN factor has the best performance in terms of MAE observed from the Removing strategy. When using the Adding strategy, the NP factor has the lowest MAE, 3.306, which means that NP factor is most closely related with the CT prediction on APS dataset. Meanwhile, the All strategy has the best performance compared with the other two strategies.

In summary, when predicting the scientific collaboration sustainability both on DBLP and APS, the CN factor is most crucial to achieving better performance, followed by the DG and NP factor. Meanwhile, the All strategy can achieve the best performance compared with the other two strategies which demonstrates the effectiveness of our selected input factors.

## 7 Conclusion

In this paper, we introduce a general model to predict scientific collaboration sustainability. First, we formulate the collaboration sustainability prediction task into two sub-questions including CD prediction and CT prediction. Then, we investigate two groups of factors including personal factors and social factors. These factors are academic ages, number of publications, number of collaborators (degree), common neighbors, and shortest path of two collaborators. Based on these input factors, we propose a novel extreme gradient boosting model named CSTeller to predict the sustainability of scientific collaboration. Extensive experimental results show that our proposed model outperforms the baseline method. The factors that determine the sustainability of scientific collaboration are analyzed with the “jackknife” approach.

Since this work is the first of its kind to study the task of collaboration sustainability prediction, there is much room for future studies in this direction. More factors may be adopted to improve the precision of this prediction task. Besides, the collaboration sustainability may result from extensive interaction among scholars. Therefore, more efforts can be done to explore the mechanism of collaboration sustainability, which may shed light on the collaboration mechanism and help policy makers to promote collaboration across institutions, disciplines, and countries.

**Acknowledgements** We thank Tong Gao for assistance with the experiments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61502071, 71774020 and 71473028, and the Fundamental Research Funds for the Central Universities under Grant (DUT18JC09).

## References

1. Adam-Bourdarios, C., Cowan, G., Germain-Renaud, C., Guyon, I., Kégl, B., Rousseau, D.: The higgs machine learning challenge. In: *Journal of Physics: Conference Series*, vol. 664, p. 072015. IOP Publishing (2015)
2. Babajide Mustapha, I., Saeed, F.: Bioactive molecule prediction using extreme gradient boosting. *Molecules* **21**(8), 983 (2016)
3. Benchettara, N., Kanawati, R., Rouveiroi, C.: Supervised machine learning applied to link prediction in bipartite social networks. In: *2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 326–330. IEEE (2010)

4. Birnholtz, J.P.: When do researchers collaborate? Toward a model of collaboration propensity. *J. Am. Soc. Inf. Sci. Technol.* **58**(14), 2226–2239 (2007)
5. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
6. Bu, Y., Ding, Y., Liang, X., Murray, D.S.: Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology* p. <https://doi.org/10.1002/asi.23966> (2017)
7. Caragea, C., Wu, J., Williams, K., Gollapalli, S.D., Khabsa, M., Teregowda, P., Giles, C.L.: Automatic identification of research articles from crawled documents. In: *WSDM 2014 Workshop on Web-scale Classification: Classifying Big Data from the Web* (2014)
8. Chakraborty, T., Patranabis, S., Goyal, P., Mukherjee, A.: On the formation of circles in co-authorship networks. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 109–118. ACM (2015)
9. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, New York (2016). <https://doi.org/10.1145/2939672.2939785>
10. Choudhury, N., Uddin, S.: Time-aware link prediction to explore network effects on temporal knowledge evolution. *Scientometrics* **108**(2), 745–776 (2016)
11. Cronin, B., Shaw, D., La Barre, K.: A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *J. Am. Soc. Inf. Sci. Technol.* **54**(9), 855–871 (2003)
12. Dong, Y., Johnson, R.A., Yang, Y., Chawla, N.V.: Collaboration signatures reveal scientific impact. In: *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 480–487. IEEE (2015)
13. Dong, Y., Johnson, R.A., Chawla, N.V.: Can scientific impact be predicted? *IEEE Trans. Big Data* **2**(1), 18–30 (2016)
14. Eom, Y.H., Jo, H.H.: Generalized friendship paradox in complex networks: The case of scientific collaboration. *Sci. Rep.* **4**, 4603 (2014)
15. Granovetter, M.S.: The strength of weak ties. *Am. J. Sociol.*, 1360–1380 (1973)
16. Hara, N., Solomon, P., Kim, S.L., Sonnenwald, D.H.: An emerging view of scientific collaboration: Scientists’ perspectives on collaboration and factors that impact collaboration. *J. Am. Soc. Inf. Sci. Technol.* **54**(10), 952–965 (2003)
17. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**(9), 1263–1284 (2009)
18. Hou, H., Kretschmer, H., Liu, Z.: The structure of scientific collaboration networks in scientometrics. *Scientometrics* **75**(2), 189–202 (2007)
19. Huang, J., Zhuang, Z., Li, J., Giles, C.L.: Collaboration over time: Characterizing and modeling network evolution. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 107–116. ACM (2008)
20. Jiang, T., Liu, T., Ge, T., Sha, L., Li, S., Chang, B., Sui, Z.: Encoding temporal information for time-aware link prediction. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2350–2354 (2016)
21. Katz, J.S., Martin, B.R.: What is research collaboration? *Res. Policy* **26**(1), 1–18 (1997)
22. Khabsa, M., Giles, C.L.: The number of scholarly documents on the public Web. *PLoS ONE* **9**(5), e93, 949 (2014)
23. Khan, S., Liu, X., Shakil, K.A., Alam, M.: A survey on scholarly data: From big data perspective. *Inf. Process. Manag.* **53**(4), 923–944 (2017)
24. Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., Tolba, A.: Exploiting publication contents and collaboration networks for collaborator recommendation. *PLoS ONE* **11**(2), e0148, 492 (2016)
25. Kong, X., Mao, M., Wang, W., Liu, J., Xu, B.: Voprec: Vector representation learning of papers with text information and structural identity for recommendation. *IEEE Transactions on Emerging Topics in Computing*. <https://doi.org/10.1109/TETC.2018.2830698> (2018)
26. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. *Science* **311**(5757), 88–90 (2006)
27. Kramer, O.: K-nearest neighbors. In: *Dimensionality Reduction with Unsupervised Nearest Neighbors*, pp. 13–23. Springer (2013)
28. Li, J., Xia, F., Wang, W., Chen, Z., Asabere, N.Y., Jiang, H.: Acrec: A co-authorship based random walk model for academic collaboration recommendation. In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 1209–1214. ACM (2014)
29. Li, L., Tong, H.: The child is father of the man: Foresee the success at the early stage. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 655–664. ACM (2015)
30. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. *J. Assoc. Inf. Sci. Technol.* **58**(7), 1019–1031 (2007)

31. Liu, H., Zhang, X., Zhang, X., Cui, Y.: Self-adapted mixture distance measure for clustering uncertain data. *Knowl.-Based Syst.* **126**, 33–47 (2017)
32. Lopes, G.R., Moro, M.M., Wives, L.K., De Oliveira, J.P.M.: Collaboration recommendation on academic social networks. In: *International Conference on Conceptual Modeling*, pp. 190–199. Springer (2010)
33. Lü, L., Zhou, T.: Link prediction in complex networks: A survey. *Physica A: Statist. Mech. Appl.* **390**(6), 1150–1170 (2011)
34. Newman, M.E.: Scientific collaboration networks. ii. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**(1), 016, 132 (2001)
35. Newman, M.E.: The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci.* **98**(2), 404–409 (2001)
36. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. *Proc. Nat. Acad. Sci.* **101**(suppl 1), 5200–5205 (2004)
37. Persson, O., Glänzel, W., Danell, R.: Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics* **60**(3), 421–432 (2004)
38. Petersen, A.M.: Quantifying the impact of weak, strong, and super ties in scientific careers. *Proc. Natl. Acad. Sci.* **112**(34), E4671–E4680 (2015)
39. Rokach, L., Maimon, O.: *Data Mining with Decision Trees: Theory and Applications*. World Scientific (2014)
40. Seber, G.A., Lee, A.J.: *Linear Regression Analysis*, vol. 936. Wiley (2012)
41. Sinatra, R., Wang, D., Deville, P., Song, C., Barabási, A.L.: Quantifying the evolution of individual scientific impact. *Science* **354**(6312), aaf5239 (2016)
42. Sonnenwald, D.H.: Scientific collaboration. *Ann. Rev. Inf. Sci. Technol.* **41**(1), 643–681 (2007)
43. Stokols, D., Hall, K.L., Taylor, B.K., Moser, R.P.: The science of team science: Overview of the field and introduction to the supplement. *Am. J. Prev. Med.* **35**(2), S77–S89 (2008)
44. Sun, Y., Han, J., Aggarwal, C.C., Chawla, N.V.: When will it happen?: Relationship prediction in heterogeneous information networks. In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pp. 663–672. ACM (2012)
45. Tang, J., Wu, S., Sun, J., Su, H.: Cross-domain collaboration recommendation. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1293. ACM (2012)
46. Tang, J., Chang, S., Aggarwal, C., Liu, H.: Negative link prediction in social media. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 87–96. ACM (2015)
47. Tsai, C.H., Lin, Y.R.: Tracing and predicting collaboration for junior scholars. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 375–380. International World Wide Web Conferences Steering Committee (2016)
48. Tylenda, T., Angelova, R., Bedathur, S.: Towards time-aware link prediction in evolving social networks. In: *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, p. 9. ACM (2009)
49. Wang, W., Bai, X., Xia, F., Bekele, T.M., Su, X., Tolba, A.: From triadic closure to conference closure: The role of academic conferences in promoting scientific collaborations. *Scientometrics* **113**(1), 177–193 (2017)
50. Wang, W., Cui, Z., Gao, T., Yu, S., Kong, X., Xia, F.: Is scientific collaboration sustainability predictable? In: *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 853–854. International World Wide Web Conferences Steering Committee (2017)
51. Williams, K., Wu, J., Choudhury, S.R., Khabsa, M., Giles, C.L.: Scholarly big data information extraction and integration in the citeseer  $\chi$  digital library. In: *2014 IEEE 30th International Conference on Data Engineering Workshops (ICDEW)*, pp. 68–73. IEEE (2014)
52. Wuchty, S., Jones, B.F., Uzzi, B.: The increasing dominance of teams in production of knowledge. *Science* **316**(5827), 1036–1039 (2007)
53. Xia, F., Chen, Z., Wang, W., Li, J., Yang, L.T.: Mvwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *IEEE Trans. Emerg. Topics Comput.* **2**(3), 364–375 (2014)
54. Xia, F., Wang, W., Bekele, T.M., Liu, H.: Big scholarly data: A survey. *IEEE Trans. Big Data* **3**(1), 18–35 (2017)
55. Yang, Z.R.: Biological applications of support vector machines. *Brief. Bioinform.* **5**(4), 328–338 (2004)
56. Zhang, C., Bu, Y., Ding, Y., Xu, J.: Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology*. <https://doi.org/10.1002/asi.23916> (2017)