Not Every Couple Is a Pair: A Supervised Approach for Lifetime Collaborator Identification

Research-in-Progress

Wei Wang

Liangtian Wan, Xiangjie Kong

Feng Xia

Zhiguo Gong

Abstract

While scientific collaboration can be critical for a scholar, some collaborator(s) can be more significant than others, a.k.a. lifetime collaborator(s). This work-in-progress aims to investigate whether it is possible to predict/identify lifetime collaborators given a junior scholar's early profile. For this purpose, we propose a supervised approach by leveraging scholars' local and network properties. Extensive experiments on DBLP digital library demonstrate that lifetime collaborators can be accurately predicted. The proposed model outperforms baseline models with various predictors. Our study may shed light on the exploration of scientific collaborations from the perspective of life-long collaboration.

Keywords: **L**ifetime collaborator, academic information retrieval, scientific collaboration

Introduction

In the past decades, while scientific data is increasingly available and interdisciplinary studies are more important (Kong *et al*. 2019, Xia *et al*. 2017), scientific collaboration plays a more important role. Many prior efforts have been done to explore the mechanism of scientific collaborations (Haines *et al*. 2011, Liu et al. 2018, Fortunato *et al*. 2018). It has been proven that authoritative scholars can be more popular in a collaboration network, and a number of collaboration recommendation algorithms/systems have been proposed based on this finding (Tang et al. 2012, Xia et al. 2014).

Scholars may experience many collaborations throughout their academic careers. However, the collaboration duration between two scholars may vary. Previous research has found that scientific collaborations are characterized by a high turnover rate juxtaposed with frequent "lifetime collaborator" (Petersen 2015). Lifetime collaborators can be more influential on scholars' academic performance. Due to the problem of academic information overload, it is not easy for scholars to find new collaborator, especially, a lifetime collaborator. Despite the importance of lifetime collaborator, there exist many open questions: Who is our lifetime collaborator? When we meet a new collaborator, can he/she become our lifetime collaborator in the future? Solving these problems can help scholars, especially junior ones, better manage and explore their academic network effectively.

Against this background, in this work-in-progress, we present a preliminary study on lifetime collaborator prediction based on the early-stage scientific collaborations. Figure 1 illustrates an

example of the proposed problem. Besides the typical structural similarity indices, we proposed a set of novel features that can enhance a life-long collaboration prediction. These features address both scholars' network properties (e.g., node degrees), and scholar-specific local properties (e.g., research interest and collaboration frequency). Through extensive experiments on the DBLP data set, we find that by integrating the proposed features, our model can achieve better performance than baseline models. Meanwhile, the contribution of each factor is explored with a jackknife approach (Dong et al. 2016). Moreover, we find that the collaboration frequency during the early collaboration stage plays a critical role in predicting lifetime collaborators.

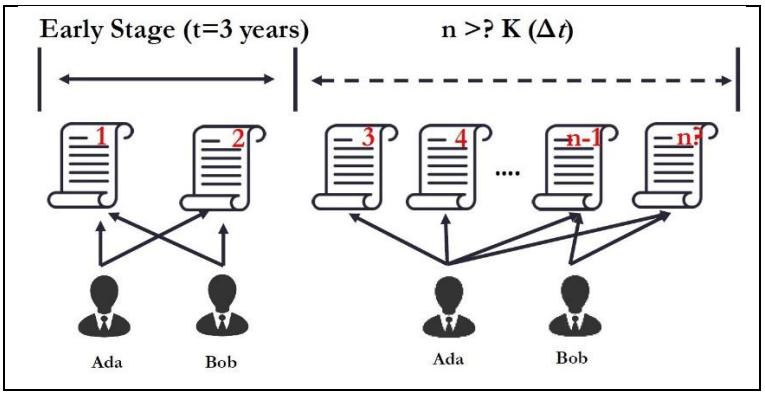

Figure 1. An example of lifetime collaborator Prediction. K is the lifetime collaborator threshold (see Equation. 1). Bob will be Ada's lifetime collaborator (n>K).

Related Works

A single scholar may not possess all the expertise or skills to tackle a complex scientific issue. Recently, scientific collaboration is becoming increasingly important. Meanwhile, interdisciplinary collaborations are far more common (Haines et al. 2011). Due to the importance of scientific collaboration, it has been extensively investigated by scholars in various disciplines including information science, social science, and computer science.

Scientific collaboration network extracted from scholarly big data is a typical way to explore the collaboration mechanism (Xia et al. 2017). Many network-based properties have been utilized to recommend scientific collaboration (Xia et al. 2014, Tsai & Lin 2016). For example, Tsai & Lin (2016) propose to predict collaboration for junior scholars based on network-based features, affiliation, geographic, and content information.

However, scientific collaborations may last for a long time which results in the phenomenon of lifetime collaborator (Petersen 2015). It has been proven that lifetime collaborator can benefit scholars in productivity and citations. Therefore, we propose to explore the phenomenon of lifetime collaborator which is a novel problem. At the same time, previous studies mainly focus on the mesoscopic features of scientific collaboration networks i.e., structural hole, or macroscopic features, while the microscopic information are overlooked (Sinatra et al. 2016). Thus, we have proposed and utilized additional local features.

The lifetime collaborator prediction can be treated as a classification task and many machine learning tools have been developed to solve such problem (Arapakis & Leiva 2016). Although many machine learning methods have achieved great success in prediction and classification, every algorithm has its own shortcomings in terms of prediction accuracy and time consumption (Fard et al. 2016). Thus, in this work, we take advantages of various machine learning approaches for lifetime collaborator prediction.

Lifetime Collaborator Prediction

We aim to develop powerful predictors to identify lifetime collaborators based on the early-stage collaborations. In this section, we introduce the proposed method in detail.

Problem Statement

We focus on the lifetime collaborator prediction problem based on the early-stage collaboration data. Given a scientific collaboration network $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, each link $\boldsymbol{e} = (\boldsymbol{i},\boldsymbol{j}) \in \mathbf{E}$ denotes collaboration relationships between scholars $\boldsymbol{i}$ and $\boldsymbol{j}$. The weight of link $\boldsymbol{w} \in \mathbf{W}$ denotes collaboration times between scholars $\boldsymbol{i}$ and $\boldsymbol{j}$. Given scholar $\boldsymbol{i}$ and his/her collaborators, our goal is to predict the whether $\boldsymbol{w_{ij}}$ will reach the lifetime collaborator threshold value $\mathbf{K}_i$ (see Eq. 1) in the future time $\boldsymbol{t'} = \boldsymbol{t} + \Delta\boldsymbol{t}$. Here, time $\boldsymbol{t}$ denotes the early collaboration stage between scholars $\boldsymbol{i}$ and his/her collaborators. In this paper, we take $\boldsymbol{t} = 3$ and $\Delta\boldsymbol{t} = 30$. In other words, we use first three-year collaboration records between scholars $\boldsymbol{i}$ and $\boldsymbol{j}$ to predict whether $\boldsymbol{j}$ is $\boldsymbol{i}$'s lifetime collaborator thirty years later. It is worth mentioning that it does not mean that a lifetime collaborator has to last at least thirty years. We set $\Delta\boldsymbol{t}$ as 30 because it is long enough to distinguish a lifetime collaborator. Then, the proposed problem is transferred to a binary classification task.

Lifetime Collaborator Definition

We define the lifetime collaborator based on the super tie in scientific collaboration network proposed by previous work (Petersen 2015), where the link weight distribution is utilized to define a super tie threshold. The scholar-specific lifetime collaborator threshold $\mathbf{K}_i$ is calculated based on the outlier statistics arguments of link weight distribution. Specifically, the threshold scholar-specific $\mathbf{K}_i$ is given by:

$$\boldsymbol{K_i} = (\langle \boldsymbol{W_i} \rangle - 1)\ln \mathbf{S}_i \tag{1}$$

where $\mathbf{S}_i$ is the number of collaborators of scholar $\boldsymbol{i}$, $\langle \mathbf{W}_i \rangle = \mathbf{S}_i^{-1} \sum_{j=1}^{Si} \boldsymbol{W_{ij}}$ is the average collaboration times between scholar $\boldsymbol{i}$ and his/her collaborator $\boldsymbol{j}$. This definition has been proven to be reasonable and effective by Peterson with the publication on PNAS (Petersen 2015). Based on the definition, the lifetime collaborator threshold $\mathbf{K}_i$ is nonparametric which depends only on the information of $\langle \mathbf{W}_i \rangle$ and $\mathbf{S}_i$. Thus, collaborators with link weight $\boldsymbol{W_{ij}} \geq \mathbf{K}_i$ are defined as the lifetime collaborators of scholar $\boldsymbol{i}$.

Input Features

We propose a series of features that may be useful to predict the lifetime collaborator. Previous research in scientific collaboration recommendation has proposed many effective features (Tsai & Lin 2016, Pecli et al. 2018). In this paper, we select four typical features as the baseline model:

- **Common Neighbors (CN)**: The CN indicates the number of common neighbors between scholars $\boldsymbol{i}$ and $\boldsymbol{j}$. To some extent, high CN means two scholars are closely related with each other in scientific collaboration networks.
- **Jaccard Coefficient (JC):** The JC measures similarity between finite sample sets. Here, the finite sample set is the co-author set.
- **Katz Weight (KW):** The KW is a node similarity measurement which considers the local path between two nodes. Here, the local paths are the collaboration network path of two given scholars.
- **Random Walk with Restart (RWR):** The RWR is a similarity index based on random walk which is an extension of PageRank algorithm. It has been proven to be effectively in recommendation systems. Note that the RWR is calculated with the largest component containing all investigated scholars extracted from the whole DBLP data set.

Although these features have been proven to be significant in link prediction problem (Tsai & Lin 2016), they may not fit the lifetime collaborator prediction. These features merely consider network

properties, while scholars' local properties are overlooked. With the availability of scholarly data we can infer more scholars' local features based on publication records. Therefore, we propose additional features, including:

- **Research Interest (RI):** The RI is proposed to measure how similar two scholars' research interests are. In order to calculate the RI, we first crawl the publication records including titles and abstracts of scholars ***i*** and ***j*** separately before they start collaborating with each other. Then, we obtain the publication corpus $\mathbf{C}_i$ (or $\mathbf{C}_j$) of scholar ***u*** (or ***v***) by integrating his/her papers together. We take advantage of the Latent Dirichlet Allocation (LDA) to $\mathbf{C}_i$ (or $\mathbf{C}_j$) to get scholar ***i*** (or ***j***)'s research topic distribution vector. Finally, we calculate the RI based on the cosine similarity of their topic distribution vectors.
- **Academic Age (AA):** The AA is proposed to describe the career stage of a given scholar. In reality, scholars have different collaboration strategies at different academic stages. For example, a PhD student will collaborate many times with his/her advisor. The student may become a colleague of his/her advisor, which may result in a life-long collaboration relationship.
- **Number of Publications (NP):** The NP is used to measure the academic achievement of a scholar based on the fact that fruitful scholars tend to be more collaborative.
- **Number of Collaborators (NC):** The NC is used to measure collaboration preference. A higher NC means that the scholar is more collaborative. Note that these above features are calculated exactly at the time when two scholars begin their collaboration.
- **Collaboration Frequency (CF):** The CF is proposed to measure how frequently two scholars collaborate with each other during the early collaboration stage. Specifically, we identify the collaboration times during the first three years of their collaboration. A higher CF may bring a stable collaboration relationship in the future.

Prediction via Machine Learning

The lifetime collaborator prediction can be treated as a binary problem. Scholar ***j*** either is the lifetime collaborator or not the lifetime collaborator of scholar ***i***. With the development of machine learning, there exist several classification methods for supervised classification. Specifically, we apply our proposed features to many advanced classification algorithms in order to predict the lifetime collaborators. These algorithms are Logistic Regression (LR), Random Forest (RF), Stochastic Gradient Decent (SGD), Support Vector Machine (SVM), Early Stage Prediction (ESP), and eXtreme Gradient Boosting (Xgboost). The ESP method is designed based on the idea in reference.

Experimental Results

Data Description

We use DBLP digital library as our research data set. Since it takes a life-long time to identify the lifetime collaborator as test set, we extract scholars whose academic age is more than 30 years as the target scholars. In other words, their first publications should be published earlier than 1986. Meanwhile, they should have no publication record in the most recent five years (from 2011 to 2016). In order to eliminate scholars who leave the academic society at their early careers (Sinatra et al. 2016), we limit our analysis to scholars who:

- have published at least one paper every five years,
- (have authored at least 10 papers,
- held their academic career for a minimum of 20 years.

Finally, as shown in Table 1, we screen out 5,631 scholars. Then, we extract all their collaborators which include 86,081 scholars. The average collaboration times (link weight **W**) is 4.125. The number of lifetime collaborators is 15,194 calculated based on Equation 1.

Table 1. Statistics of experimental data set

Scholars	Collaborators1	Lifetime Collaborators	<W>
5,631	86,081	15,194	4.125

Experimental Design

To predict lifetime collaborators, we divide the dataset into two non-overlapping subsets. The first sub set is the training set and the second is the test set. The test set is 20% of the whole dataset. Meanwhile, the k-fold (k=5) cross validation is adopted in the experiments to enhance the stability and fidelity of evaluation results. For a given scholar, we divided his/her collaborators into two kinds, i.e., lifetime collaborators (positive samples) and others (negative samples). It is worth mentioning that the positive and negative samples are unbalanced. For example, if scholar ***i*** has ten collaborators, only two scholars are the positive samples (lifetime collaborator) and the rest eight scholars are negative samples (none-lifetime collaborator). All the input features are normalized into [0, 1] in order to avoid the data imbalance with the min-max normalization approach. The prediction results are evaluated with four typical metrics including Accuracy, Precision, Recall, and F1. 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.

Predictor Comparison

Table 2. Comparison of baseline model and proposed model with various predictors in terms of accuracy, precision, recall, and F1.

Predictor	Features	Accuracy	Precision	Recall	F1
LR	Baseline	0.8106	0.7862	0.5816	0.6203
	All_F	0.8016	0.7816	0.6134	0.6152
RF	Baseline	0.8743	0.7425	0.6543	0.6524
	All_F	0.8842	0.7318	0.6734	0.6354
SGD	Baseline	0.8213	0.7689	0.6126	0.6524
	All_F	0.8331	0.7858	0.6515	0.6852
SVM	Baseline	0.9022	0.7851	0.6563	0.6583
	All_F	0.9112	0.7828	0.6815	0.6701
ESP	Baseline	0.8854	0.7888	0.6675	0.5927
	All_F	0.9013	0.7998	0.6901	0.6234
Xgboost	Baseline	0.9106	0.7922	0.6616	0.6503
	All_F	0.9216	0.8165	0.6874	0.6852

Table 2 shows the performance of different predictors on DBLP data set. In the table, the baseline model takes CN, JC, KW, and RWR as the input features. As can be seen from this table, all predictors have at least 80% accuracy, 73% precision, 57% recall, and 61% F1. This indicates that the lifetime collaborator can be well predicted with our proposed model. The Xgboost predictor achieves the highest accuracy (91.06%) and precision (79.02%). However, the recall rate for all predictors

merely ranges from 57.34% to 67.34%, which indicates that the baseline model is not sensitive enough to predict all the possible lifetime collaborators.

The All_F model takes all baseline feature as well as RI, AA, NP, NC, CF as input features. By comparison, we can see that the All_F model could achieve better performance than baseline model, which demonstrates the effectiveness of our proposed features. Meanwhile, the Xgboost still achieves the best performance with accuracy 92.16% and 81.65% precision. From Table 2, we can gain the conclusion that lifetime collaborators can be well predicted with our proposed model. The proposed features can improve the performance of lifetime collaborator prediction than baseline model.

Factor Contribution Analysis

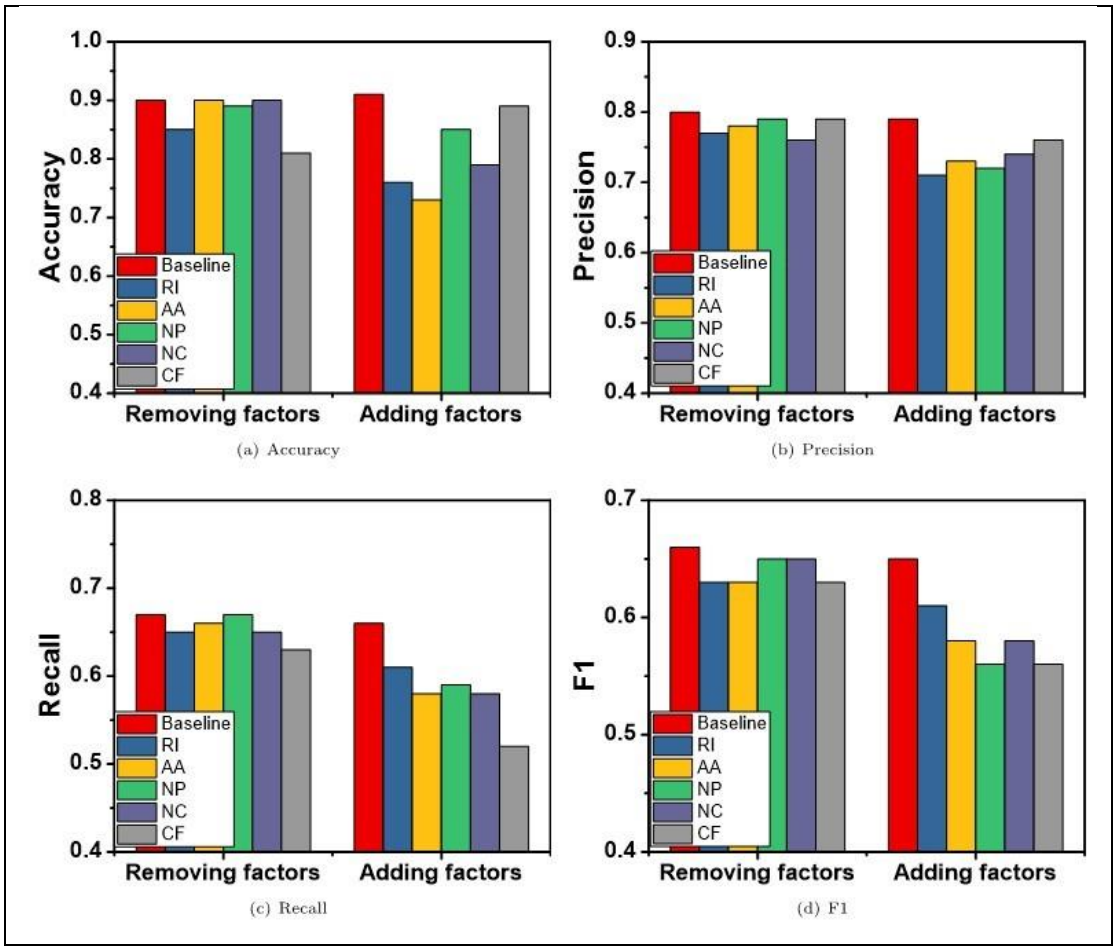

Figure 2. Feature contribution comparison in terms of accuracy, precision, recall, and F1 with adding or removing strategies. The left and right sides of the figure depict the effects of removing strategy and adding strategy, respectively. The baseline feature includes CN, JC, KW, and RWR.

We have utilized a total of six features, including the baseline model, as follows: baseline (CN, JC, KW, and RWR), RI, AA, NO, and CF that may determine a life-long collaboration relationship. In order to reveal the significance of each proposed feature, we employ the "jackknife" approach (Dong et al. 2016) with two cases:

- Removing one factor and predicting with the rest factors (Removing);
- Using only one factor to do prediction (Adding).

Based on this approach, we can explore the individual contribution that each feature supports to the overall prediction problem. At the same time, we utilize the Xgboost as the predictor in this experiment.

Figure 2 shows the results of accuracy, precision, recall, and F1 with jackknife approach based on Xgboost. We can find that our proposed All F model achieves better performance than the adding and removing factor strategies in most cases. In subfigure 2 (a), 16% drops in accuracy (from 0.92 to 0.76) by the removal of CF feature. Meanwhile, the accuracy of adding strategy with CF feature almost remains unchanged. These indicates that the CF feature has significant impact on lifetime collaborator prediction. The experiment also shows that the proposed novel features can be important to enhance the prediction precision, recall and F1.

Conclusion

In this work, we try to predict the lifetime collaborator of a given scholar based on the early-stage collaboration data. In order to solve this novel problem, we propose a number of features that may be useful to predict a life-long collaboration relationship. Through extensive experiments on the DBLP data set, we find the model by leveraging the proposed novel features outperform the baseline model. This finding is also consistent by using factor analysis via jackknife approach.

In future work, we will investigate more sophisticated graphical and local features while proposing novel learning model to further enhance the predication performance. Meanwhile, we would like to take advantages of network representation learning methods to acquire scholar vectors for similarity calculation.

Reference

Arapakis, I. & Leiva, L. A. (2016), Predicting user engagement with direct displays using mouse cursor information, in 'Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval', ACM, pp. 599–608.

Dong, Y., Johnson, R. A. & Chawla, N. V. (2016), 'Can scientific impact be predicted?', IEEE Transactions on Big Data 2(1), 18–30.

Fard, M. J., Wang, P., Chawla, S. & Reddy, C. K. (2016), 'A bayesian perspective on early stage event prediction in longitudinal data', IEEE Transactions on Knowledge and Data Engineering 28(12), 3126–3139.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojevi´c, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B. et al. (2018), 'Science of science', Science 359(6379), eaao0185.

Haines, V. A., Godley, J. & Hawe, P. (2011), 'Understanding interdisciplinary collaborations as social networks', American journal of community psychology 47(1-2), 1–11.

Kong, X., Shi, Y., Yu, S., Liu, J., & Xia, F. (2019), 'Academic social networks: Modeling, analysis, mining and applications', Journal of Network and Computer Applications, 132(1), 86-103.

Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q. & Lee, I. (2018), 'Artificial intelligence in the 21st century', IEEE Access 6, 34403–34421.

Pecli, A., Cavalcanti, M. C. & Goldschmidt, R. (2018), 'Automatic feature selection for supervised learning in link prediction applications: a comparative study', Knowledge and Information Systems 56(1), 85–121.

Petersen, A. M. (2015), 'Quantifying the impact of weak, strong, and super ties in scientific careers', Proceedings of the National Academy of Sciences 112(34), E4671–E4680.

Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. (2016), 'Quantifying the evolution of individual scientific impact', Science 354(6312), aaf5239.

Tang, J., Wu, S., Sun, J. & Su, H. (2012), Cross-domain collaboration recommendation, in 'Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1285–1293.

Tsai, C. H. & Lin, Y. R. (2016), Tracing and predicting collaboration for junior scholars, in 'Proceedings of the 25th International Conference Companion on World Wide Web', International World Wide Web Conferences Steering Committee, pp. 375–380.

Xia, F., Chen, Z., Wang, W., Li, J. & Yang, L. T. (2014), 'Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors', IEEE Transactions on Emerging Topics in Computing 2(3), 364–375.

Xia, F., Wang, W., Bekele, T. M., & Liu. H. (2017), 'Big scholarly data: A survey', IEEE Transactions on Big Data 3(1), 18-35.