Is scientific collaboration sustainability predictable?

Wei Wang, Zixin Cui, Tong Gao, Jiaying Liu, Xiangjie Kong, Feng Xia
School of Software, Dalian University of Technology, Dalian 116620, China

Abstract
This work aims to explore whether the sustainability of scientific collaboration including collaboration duration and collaboration times can be predicted. For this purpose, we employ a linear regression model. The features used in the linear regression model contain the factors extracted from the local characteristics and network properties of scholars. We take advantage of the open data from DBLP to evaluate the extent to which the sustainability of scientific collaboration can be predicted. To the best of our knowledge, this is the first work that tries to investigate the collaboration dynamics from the perspective of sustainability. Our findings on the sustainability of scientific collaboration can shed light on the problems of link predictions and collaborator recommendations.

Keywords: Scientific collaboration; liner regression; sustainable collaboration
DOI: Citation info is to be added.

Contact: f.xia@acm.org.

1 Introduction

Scientific collaboration plays an important role in scientific research. Previous research has demonstrated that collaboration will greatly influence the publication quality and quantity, and scholars are becoming more and more collaborative (Wang et al., 2016). A lot of efforts have been done to reveal the mechanisms of scientific collaborations. For example, the link prediction (Lü & Zhou, 2011) and collaborator recommendation (Xia, Chen, Wang, Li, & Yang, 2014) in scientific information networks have been extensively analyzed. However, previous studies mainly focus on whether two scholars will collaborate (i.e., new links emerge between two nodes) in the future based on the similarity or network structure. Few works have been done on collaboration mechanisms after the collaboration has been established. As shown in Figure 1, after the first collaboration, scholars may choose to continue or terminate collaborating. Meanwhile, the collaboration duration between two scholars is uncertain.

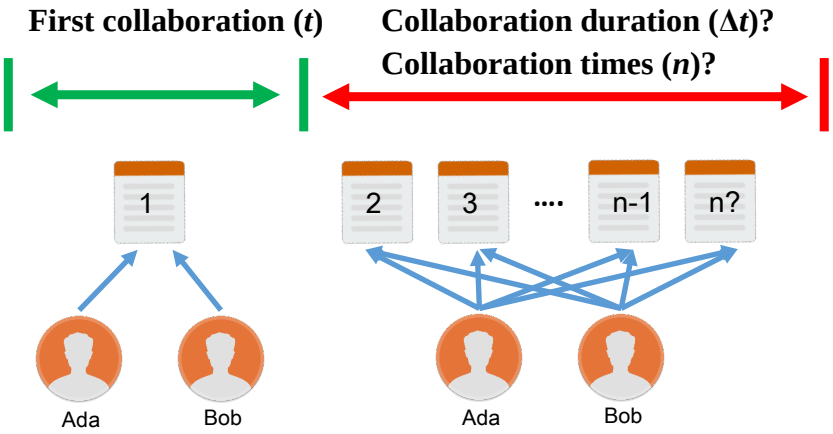

Figure 1: An example of scientific collaboration sustainability.

In this paper, we formulate these tasks as the sustainability of scientific collaborations. Our proposed problem of collaboration sustainability prediction is fundamentally different from the traditional problems

of the link prediction and collaborator recommendation. We try to explore the factors that determine the collaboration sustainability from the local and network properties of scholars. Moreover, we employ a linear regression model to predict the collaboration sustainability after first-time collaboration. Trough experiments on DBLP data set, we find that the sustainability of scientific collaboration can be well predicted with the factors we use.

2 Scientific Collaboration Sustainability Prediction

2.1 Experimental Setup

Our primary task is to predict how long and how many times two scholars will collaborate after timestamp t which represents the time of first collaboration. As shown in Figure 1, we try to predict the time duration Δt and the number of collaboration n (collaboration times) in Δt.

2.2 Data Sets

We extract scholars who have more than 10 papers from 1950 to 2016 in DBLP as our data sets which contain $185,739$ scholars and $1,721,923$ collaboration records. We then crawl all factors needed from DBLP. We randomly select 80% nodes as training set and the rest as the testing set.

2.3 Methods

Inspired by previous work by Dong et al. (Dong, Johnson, & Chawla, 2016), we formulate the collaboration sustainability prediction problem as a regression task. We employ the linear regression (Seber & Lee, 2012) as our basic model because of its effectiveness and convenience. Considering predicting the collaboration sustainability between scholar A and B, the features we use are shown in Table 1. Specifically, factors N, D, and A are scholars' local characteristics, and factors C and S are scholars' network properties.

Factors	Description
Number of papers (N)	the number of publications of A and B before collaboration
Degree (D)	the number of collaborators of A and B before collaboration
Academic Age (A)	the academic ages of A and B when first collaborating
Common Neighbors (C)	the number of common neighbors of A and B before collaboration
Shortest Path (S)	the shortest path between A and B before collaboration

Table 1: Factor definitions

2.3.1 Linear Regression

If the collaboration times x (or duration) are described by d features, we can define the description as $x = (x_1; x_2; ...; x_d)$. The liner regression tries to find a function $f(x)$ for prediction, which is described as:

$$f(x) = \omega_1 x_1 + \omega_2 x_2 + ... + \omega_d x_d + b \tag{1}$$

or in vector form as:

$$f(x) = \omega^T + b \tag{2}$$

where, ω and b are learned from the training sets.

2.3.2 Evaluation Metrics

We use the mean absolute error (MAE) to evaluate the performance of prediction model, which can be calculated as:

$$E(f; D) = \frac{1}{d}\sum_{i=1}^{d}(f(x_i) - y_i)^2 \tag{3}$$

3 Results

To investigate the contributions and influence of each input factor on the prediction task, we employ a "jackknife" approach (Dong et al., 2016) with three cases: (1) we remove one factor and calculate the MAE only with the rest factors (Removing factors); and (2) we use only one factor to calculate the MAE (Adding factors); and (3) we calculate the MAE with all factors (All factors). Figure 2 reports the predictive results of the collaboration duration in term of MAE. On the one hand, the strategy of removing factors has a relative lower MAE than that of Adding factors, which means that the factors we consider do improve the performance of prediction. On the other hand, the method considering all factors has the best results. Figure 3 illustrates the results of collaboration times prediction. The overall trend in Figure 3 is similar with the trend in Figure 2. The experimental results in these two figures indicate that the collaboration sustainability can be predicted and it is effective to employ the local characteristics and network properties as input features .

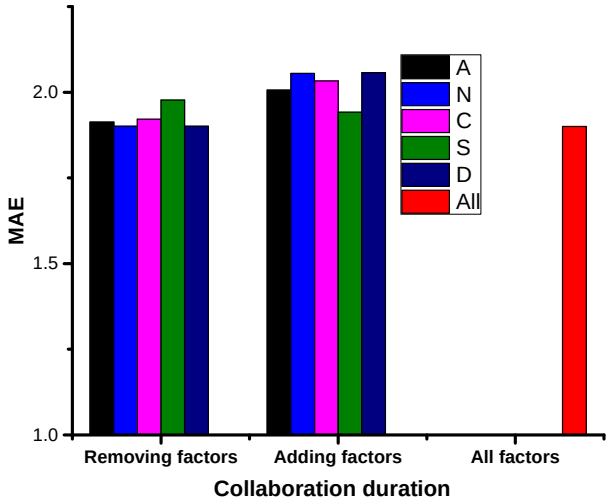

Figure 2: MAE of the duration prediction.

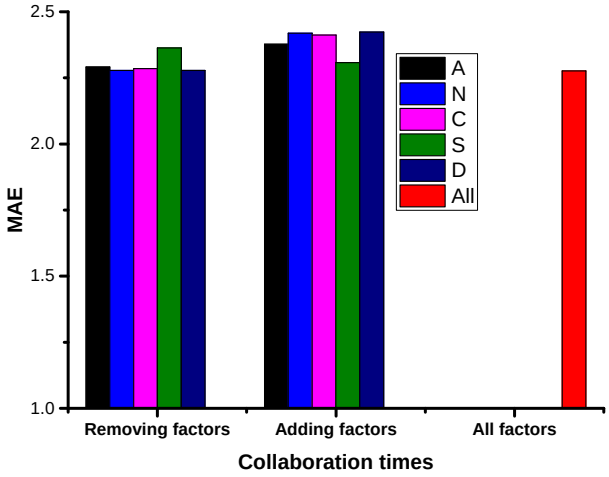

Figure 3: MAE of the times prediction.

4 Conclusion

In this work, we have proposed to predict the sustainability of scientific collaboration with a linear regression model. We consider both the local characteristics and network properties of scholars as input features. Experimental results show that the factors we employ are effective in predicting the sustainability of scientific collaboration. In the future, we will try other predictive models and more input features to better understand the sustainability of scientific collaborations.

References

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

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, *390*(6), 1150–1170.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 936). John Wiley & Sons.

Wang, W., Liu, J., Yu, S., Zhang, C., Xu, Z., & Xia, F. (2016). Mining advisor-advisee relationships in scholarly big data: A deep learning approach. In *Proceedings of the 16th acm/ieee-cs on joint conference on digital libraries* (pp. 209–210).

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.