Is Scientific Collaboration Sustainability Predictable?

Wei Wang, Zixin Cui, Tong Gao, Shuo Yu, Xiangjie Kong and Feng Xia
School of Software, Dalian University of Technology
Dalian 116620, China
xjkong@acm.org

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 propose a series of features including structural similarity indices, authorship properties, and research interests. Experimental results on a real-world dataset show that our proposed model outperforms baseline model by 10% in MAE. Our study may shed light on investigating scientific collaboration from the perspective of sustainability.

Keywords

Scientific collaboration; Sustainable collaboration; Early stage prediction

1. INTRODUCTION

Scientific collaboration plays an important role in scientific research. Numerous efforts have been done to reveal the mechanisms of scientific collaborations. For example, link prediction [1] in scientific information networks has 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. Few attention has been paid to explore the mechanisms after the collaboration has been established. In reality, scientific collaboration is not a one-shot deal. As shown in Figure 1, after the first collaboration, scholars may collaborate with each other more than once. In other words, scientific collaborations are sustainable.

In this work-in-progress paper, we try to explore the mechanism of sustainable collaboration and predict the sustainability of scientific collaboration accordingly. We try to answer the following questions as long as the collaboration is established: (i) How long will it last? (ii) How many times will they collaborate? (iii) Can the sustainability of this collaboration be predicted? We propose a series of features that may determine the sustainability of scientific collaboration and predict it from the perspectives of collaboration duration and collaboration times. Experimental results on a real-world dataset with five predictors demonstrate the effectiveness of our proposed features.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3041021.3054230

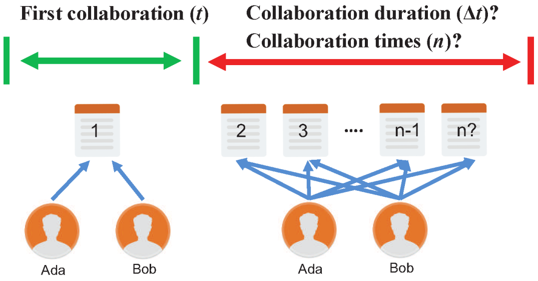
First collaboration (*t*)
Collaboration duration (Δ*t*)?
Collaboration times (*n*)?
1
2
3
....
n-1
n?
Ada
Bob
Ada
Bob

Figure 1: An example of sustainable collaboration.

2. SUSTAINABILITY PREDICTION

2.1 Problem Statement

We formulate the scientific collaboration sustainability prediction as two prediction problems including Collaboration Duration (CD) prediction and Collaboration Times (CT) prediction. Given a social network $G = (V, E)$, where each edge $e = (u, v) \in E$ denotes the coauthorship between scholars u and v when they first collaborate with each other at time t. Our goal is to predict the future duration and times of coauthorship in the future time $t' > t$ (i.e., $t' = t+20$). In other words, the goal is to predict how long and how many times will two scholars collaborate with each other in the future t' based on the data observed when they begin their collaboration. The prediction problem is a regression task.

2.2 Data Description

To eliminate scholars who leave research at their early career, we limit our analysis to scientists who: (i) have published at least one paper every 5 years, (ii) have authored at least 10 papers, (iii) their publication career spans at least 20 years. In order to predict the sustainability with sufficient time span, we further limit our study to papers published up to 2000. The CD and CT are calculated in the following 20 years after the first-time collaboration. Finally, we screen out 14,022 scholars and their 355,992 collaborators from the DBLP Digital Library. The average <CD> and <CT> are 3.07 and 3.82, respectively.

2.3 Feature Extraction

We first select five structural similarity features as our baseline model, including Common Neighbors (CN), Jaccard Coefficient (JC), Adamic-Adar (AD), Katz_Weight (KW), and Preferential Attachment (PA). These features have been proven to be effective in prior work for missing link prediction [2]. However, these features may not have good performance in sustainability prediction because they merely consider network proximity. Therefore, we propose additional features, including:

Shortest Path (SP): The SP is to measure how close two scholars are in the scientific collaboration network. In order to calculate the SP, we use all the publications of entire DBLP. For example, if scholars u and v begin their collaboration at 2000, we construct the scientific collaboration network G_{uv} by extracting all the publications before 2000.

Research Interest (RI): The RI is to measure the similarity of two scholars' research interests. We first crawl the paper information including titles and abstracts of scholars u and v separately before they start collaborating with each other. Then, we get the research corpus C_u (or C_v) of scholar u (or v) by integrating his/her publication record together. We apply the Latent Dirichlet Allocation (LDA) to C_u (or C_v) to get scholar u (or v)'s topic distribution vector. Finally, the RI is calculated based on the cosine similarity of two scholar's topic distribution vectors.

Academic Age (AA): The AA is to describe a scholar's career stage. We take advantages of the AA feature based on the fact that scholars tend to have different collaboration strategies at different academic stages. For example, a PhD student will collaborate frequently with his/her advisor.

Number of Publications (NP): The NP is used to measure the academic performance of each scholar based on the idea that fruitful scholars tend to be more collaborative and have a higher reputation.

Number of Collaborators (NC): The NC is used to measure how collaborative a scholar is. Note that all these features are calculated exactly the time when two scholars begin their collaboration.

2.4 Prediction via Regression

There are many regression algorithms for collaboration sustainability prediction. All proposed features are applied to five predictors including Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Xgboost, and Early Stage Prediction (ESP). The ESP algorithm is designed following the idea in [3]. Rest algorithms are implemented in Python with related packages.

3. EXPERIMENT RESULTS

We divide the whole dataset into two subsets including training set and testing set. We randomly select 80% dataset as the training set and the rest as the testing set. All the input features are normalized into $[0, 1]$ in order to avoid the data imbalance with the min-max normalization. All the results are calculated by the average values of k-fold (k=10) cross validation. The evaluation metrics are MAE (Mean Absolute Error), MSE (Mean Square Error), PCC (Pearson's Correlation Coefficient), and CCC (Concordance Correlation Coefficient).

Tables 1 and 2 show the performance of different predictors on CD prediction and CT prediction, respectively. The comparison of each predictor between baseline features (Baseline) and all proposed features (All_F) is illustrated. We can observe that all predictors can achieve good performance in both CD and CT prediction, which demonstrates the feasibility of sustainability prediction. The results show that Xgboost is the best predictor. Meanwhile, by adding new features, all predictors achieve better performance compared with baseline features, which demonstrates the effectiveness of these features. For example, Xgboost with A_F is 10% and 12% lower in MAE in terms of CD and CT prediction, respectively. We can also see that CD can be better predicted than CT. The reason is that scholars' CT is always higher than CD. In other words, scholars may collaborate more than once in one year.

Table 1: Results of CD prediction

Predictor	Features	MAE	MSE	PCC	CCC
LR	Baseline	1.71	8.09	0.61	0.51
	All_F	1.68	7.85	0.64	0.52
RF	Baseline	1.43	7.45	0.68	0.54
	All_F	1.42	7.38	0.70	0.54
SVM	Baseline	1.32	6.89	0.66	0.54
	All_F	1.31	6.58	0.65	0.52
Xgboost	Baseline	1.22	6.51	0.73	0.58
	All_F	1.11	6.28	0.75	0.60
ESP	Baseline	1.35	6.24	0.65	0.57
	All_F	1.31	6.20	0.70	0.62

Table 2: Results of CT prediction

Predictor	Features	MAE	MSE	PCC	CCC
LR	Baseline	2.31	12.39	0.51	0.46
	All_F	2.28	11.56	0.51	0.48
RF	Baseline	2.33	13.05	0.58	0.47
	All_F	2.30	13.18	0.61	0.48
SVM	Baseline	2.13	11.89	0.59	0.49
	All_F	2.11	16.07	0.60	0.50
Xgboost	Baseline	1.78	10.51	0.61	0.51
	All_F	1.65	9.23	0.62	0.51
ESP	Baseline	1.85	10.24	0.62	0.52
	All_F	1.81	9.81	0.62	0.52

4. CONCLUSIONS

Scientific collaboration is sustainable, where two scholars may collaborate for many times. In this paper, we propose to predict the sustainability of scientific collaboration as long as two scholars begin their collaboration. This proposed problem is fundamentally different from the traditional problems including link prediction and collaborator recommendation. Our research may shed light on exploring the mechanisms of scientific collaborations.

5. REFERENCES

[1] C. H. Tsai and Y. R. Lin. Tracing and predicting collaboration for junior scholars. In *Proc. of the 25th WWW*, pages 375–380. Intl. WWW Conferences Steering Committee, 2016.

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

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