Network analysis of scientific collaboration

Elena Novikova
Insitute of Computer Science
University of Tartu

Tartu, Estonia
elena.novikova@ut.ee

https://github.com/ElenaSNovikova/NS_Network_analysis_of_scientific_collaboration

Abstract – Interdisciplinary collaboration is often a key for innovation and solving complex problems. On the other hand, collaboration can considerably boost a researcher's career. Trying to minimize the chance of dissolution of the collaboration can also help to get a deeper understanding of the evolution of complex networks. That's why collaboration recommendation has been studied so much lately. Unfortunately, long-term collaboration, which is associated with a better publication history, has received little attention. In this paper, supervised link prediction (binary classification) was used together with the ArnetMiner data to construct and analyse a collaboration networks of Computer Science researchers. In the case where only two classifiers (Logistic Regression and Random Forest Classifier) were used, it was found that the suitability of feature selection approach doesn't depend on the choice of the classifier. The results show graph embedding with Node2Vec to be faster, but less accurate than using scoring measures as features.

Keywords— collaboration network, link prediction

I. INTRODUCTION

Interdisciplinary collaboration is increasingly important for innovation and finding possible solutions for complex problems. In this era of global health crisis interdisciplinary research could also plays a crucial role. Insights from diverse perspectives could provide a clearer understanding and new ideas.

Scientists are often faced with the important decision to start or terminate a collaboration. Yet little attention has been paid to the finding long-term collaborators, although collaborations and scientific exchange will often lead to continuous improvement. A recent study of the life-long careers of 3,860 computer scientists found that 72.7% of coauthors are transient, i.e. collaborating once but never again with the same scientists. Another study looked at 473 physicists and biologists over the course of their careers, which covered 94,000 papers and involved 166,000 collaborators. Between 60 and 80 percent of any given researchers' collaborations lasted just one year or even less. It was found that long-term collaboration increased publication numbers for an individual researcher by, on average, 17 percent, and in so doing boost his or her career development and success.

The study of scientific collaborations could take various direction. Researchers and papers can be considered as two distinct groups of nodes, where a researcher is linked to all papers that he or she has co-authored (bipartite collaboration networks). In the case of paper networks, papers are considered as nodes that are connected if they have a common author. In the case of collaboration networks, researchers are considered as nodes that are connected if they have co-authored at least one paper. In this project, such collaboration network was constructed, which is considered an undirected weighted network G (V, E) at a particular time slice, where V and E are sets of nodes and links, respectively.

In this project two different approaches were studied: constructing scoring methods and implementing Node2Vec. Two different classifiers were employed: Logistic Regression and Random Forest Classifier. On the AMiner-Author dataset scoring methods approach is superior and Logistic Regression shows higher roc_auc_score.

Link prediction is an important issue for scientific collaboration networks. Both the formation and dissolution of links could vividly illustrate the existing network dynamics. While studying the process of link creation the network evolution could be predicted. While analysis of link dissolution could shed the light on the network evolution from completely different point of view.

II. RELATED WORK

In this section, some of the existing works on collaboration networks are presented. As it is a well-studied field there are various direction that the study of such networks could take.

A. Collaboration networks and Weighted collaboration networks

While constructing a network researchers are considered as nodes, that are connected if they have co-authored at least one paper. The focus of study could be (1) different fields of research, regions or combinations of thereof; (2) the impact of scientific collaboration; (3) publication habits.

To take the strength of a collaboration into account the weighted network could be constructed. The number of collaborations between researchers would be the weight of the link in the previously described network. In [3] such network is used to find academic-age-aware collaboration patterns.

Centrality measures, community structure detection techniques and link prediction techniques could be applied to such networks. In [2] a new link prediction technique is proposed to identify possible interdisciplinary collaboration. [4] is focused on scientific communities and how they evolve over time.

B. Bipartite collaboration networks

While constructing a network researchers and papers are considered as two distinct group of nodes, where a researcher is linked to all papers, that he or she has co-authored. [6] focuses on extracting the backbone from bipartite projections and overcoming their limitations.

C. Collaboration hypergraphs

While constructing a network researchers are considered as hypergraph nodes, that are connected by hyperedges, representing papers. Another approach is to consider papers as hypergraph nodes, that are connected by hyperedges representing researchers. In [1] such approach is implemented to measure an influence of a researcher over collaboration.

D. *Papers network*

While constructing a network papers are considered as nodes, that are connected if they have a common author. [5] shows the benefits of such a network: a better defined community structure and better represented results of collaborations – the published papers.

III. DATASET

In this section, datasets necessary for completing this project are presented. After giving the description and the source of the datasets, data cleaning and data preparation are described. the results of the descriptive analysis are presented as well.

A. *Description*

To construct a collaboration network, I used the data from ArnetMiner (https://www.aminer.cn/aminernetwork) [8]. It includes the data on scientific papers, authors and their collaboration.

The dataset AMiner-Author includes the detailed information about authors.

TABLE I. DATASET AMINER-AUTHOR

Columns	Description
Index	Index of the author
Author	Name of the author
Affiliation	Affiliation of the author
Papers	Count of published papers of the author
Citations	Total number of citations of the author
H_id	H-index of the author
P_id	P-index with equal A-index of the author
uP_id	P-index with unequal A-index of the author
Research	Research interests of the author

Fig. 1. Description of the dataset AMiner-Autho

The dataset AMiner-Coauthor includes the index of the first author, the index of the second author and the number of collaborations between them.

B. *Data Preparation*

In the dataset AMiner-Coauthor there was initially 4258946 collaborations and no missing values. There were no duplicated rows. There were no rows for which "target" and "source" were the same.

In the dataset AMiner-Author there was initially 1712433 authors. There were no duplicated rows. The only missing values were in the column "Affiliation", that were filled with value "NaN".

After using the dataset AMiner-Coauthor I created an undirected weighted network with 1560640 nodes, 4258946 edges and 156240 connected components. Because of the high computational cost I decided to only analyse scientists that focus on "Computer Sciece" (the variable "Research" from the dataset AMiner-Author). To get a connected network I have also chosen the largest subgraph from this network.

C. *Descriptive analysis*

The constructed undirected weighted network has 1707 nodes, 4923 edges and average degree of 5.7680. From the degree histogram below we can observe that the majority of nodes has degree of 1.

PLOT I. DEGREE HISTOGRAM

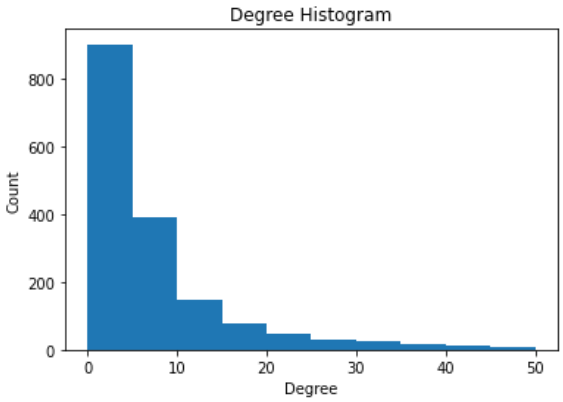

Fig. 2. Degree Histogram for undirected weighted collaboration network

It should be noted that a really low fraction of nodes is connected on average (clustering coefficient of 0.0294) and that the ratio of existing edges over possible ones is pretty low as well (edge density of 0.0034). The longest distance between two nodes is equal to 13, while average path length is equal to 6.3849.

In this project various centrality measures were calculated as well.

TABLE II. CENTRALITY MEASURES

Degree centrality	Robert David McCartney Kathryn E. Sanders Mark Guzdial Beth Simon Lynda Thomas
Closeness centrality	John Impagliazzo Henry M Walker Deborah L. Knox Mark Guzdial Owen Astrachan
Eigenvector centrality	Robert David McCartney Kathryn E. Sanders Carol Sue Zander Lynda Thomas Anna Eckerdal
Betweenness centrality	N. B. Dale Owen Astrachan Mark Guzdial Ursula Wolz Joyce Currie Little
Pagerank	Mark Guzdial Gerald L. Engel Susan Rodger N. B. Dale Robert David McCartney

Fig. 3. Top 5 scientists according to various centrality measures

According to these results, Robert David McCartney and Mark Guzdial deserve special attention.

IV. METHODOLOGY

After constructing a weighted collaboration network and conducting the descriptive analysis I focus on the problem of link prediction. Link prediction could be used to mine and analyse the evolution of networks. It should be noted that the link prediction is a binary classification problem.

In this project I follow the methodology that we focused on during practice sessions. The first approach focused on calculating various scoring methods, that later could be used in supervised machine learning techniques. To take into account the increasing calculation costs I focused on:

- Neighborhood measures (Number of common neighbors, Jaccard coefficient, Adamic/Adar)
- Path-based measures (Shortest path, Hitting time)
- Vertex feature aggregation (Preferential attachment)

After obtaining these features I employed Logistic Regression and Random Forest Classifier. To efficiently compare them I calculated various metrics: roc auc score, accuracy score, f1 score, recall score, precision score and mean squared error.

The second approach focused on implementing Node2Vec. After calculating the embeddings Logistic Regression and Random Forest Classifier are once again employed and various metrics are calculated.

V. RESULTS

While constructing the data to solve the data link prediction problem, I used the method we followed during practice sessions. Unfortunately, it leads to highly imbalanced data. As a result, some classifier could always predict the absence of link, yet at the same time give very high accuracy score and zero precision score.

As a result, while choosing the better approach I used roc_auc_score. While comparing Logistic Regression and Random Forest Classifier I checked other metrics as well.

After implementing both approaches the following results were obtained.

TABLE III. SCORING METHODS AND NODE2VEC COMPARISON

	Scoring Methods roc_auc_score	Node2Vec roc_auc_score
Logistic Regression	0.6686	0.6035
Random Forest Classifier	0.6675	0.6298

Fig. 4. roc auc scores for both approaches

As scoring methods show similar results, ROC Curve Comparison was plotted.

PLOT II. ROC CURVE COMPARISON

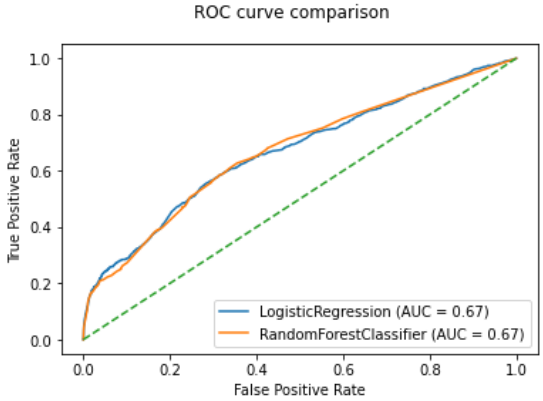

Fig. 5. ROC Curves for scoring method approach

As scoring methods approach has shown better results some other metrics were also included in this report.

TABLE IV. SCORING METHODS FOR LOGISTICS REGRESSION

	Logistic Regression
roc auc score	0.6686
accuracy score	0.7276
f1 score	0.2301
recall score	0.5092
precision score	0.1487
mean squared error	0.2724

Fig. 6. Various metrics calculated for Logistic Regression

VI. CONCLUSIONS

In this paper, an undirected weighted collaboration network of Computer Science researchers was constructed from the Aminer datasets. The descriptive analysis showed that scientists often collaborate only once; low fraction of authors is connected on average. The ratio of existing collaborations over possible collaborations is quite low as well.

Two classifiers, Logistic Regression and Random Forest Classifier, and two approaches for feature selection, graph embedding and scoring methods, were used for link prediction. Although the Node2Vec approach was faster, scoring methods approach allowed for higher accuracy in both cases. So, it was found that the choice of feature selection approach doesn't depend on the choice of the classifier. The best combination was Random Forest Classifier and scoring methods as features. The computational cost was a constant problem.

There are various possible directions for the future research. One of them could be the problem of Link dissolution. Another direction could be prediction of long-term collaborations. Possible solution to reduce high computational complexity could be the sliding window algorithm.

REFERENCES

[1] Lung, R.I., Gaskó, N. & Suciu, M.A. A hypergraph model for representing scientific output. Scientometrics 117, 1361–1379 (2018). https://doi.org/10.1007/s11192-018-2908-2

[2] Cho, H., Yu, Y. Link prediction for interdisciplinary collaboration via co-authorship network. Soc. Netw. Anal. Min. 8, 25 (2018). https://doi.org/10.1007/s13278-018-0501-6

[3] Wang, W., Yu, S., Bekele, T.M. et al. Scientific collaboration patterns vary with scholars' academic ages. Scientometrics 112, 329–343 (2017). https://doi.org/10.1007/s11192-017-2388-9

[4] B. Grba and A. Meštrović, Tracking the evolution of scientific collaboration networks. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0503-0508, doi: 10.23919/MIPRO.2018.8400096.

[5] Gaskó, N., Lung, R.I. & Suciu, M.A. A new network model for the study of scientific collaborations: Romanian computer science and mathematics co-authorship networks. Scientometrics 108, 613–632 (2016). https://doi.org/10.1007/s11192-016-1968-4

[6] Neal, Z. (2014). The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. Social Networks, 39, 84–97.

[7] B. Xu, L. Li, J. Liu, L. Wan, X. Kong and F. Xia, "Disappearing Link Prediction in Scientific Collaboration Networks," in IEEE Access, vol. 6, pp. 69702-69712, 2018, doi: 10.1109/ACCESS.2018.2880233.

[8] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

[9] Jun Wen, Wei Wang, Metin Kozak, Xinyi Liu & Haifeng Hou (2020) Many brains are better than one: the importance of interdisciplinary studies on COVID-19 in and beyond tourism, Tourism Recreation Research, DOI: 10.1080/02508281.2020.1761120

[10] Petersen, Alexander Michael. "Quantifying the impact of weak, strong, and super ties in scientific careers." Proceedings of the National Academy of Sciences 112.34 (2015): E4671-E4680.

[11] Cabanac, G., Hubert, G., Milard, B. (2015). Academic careers in Computer Science: Continuance and transience of lifetime co-authorships. Scientometrics, 102(1), 135–150