# Scientific Matchmaker: Collaborator Recommender System

Ilya Makarov(✉), Oleg Bulanov, Olga Gerasimova,
Natalia Meshcheryakova, Ilia Karpov, and Leonid E. Zhukov

National Research University Higher School of Economics,
Kochnovskiy Proezd 3, 125319 Moscow, Russia
iamakarov@hse.ru, revan1986@mail.ru

**Abstract.** Modern co-authorship networks contain hidden patterns of researchers interaction and publishing activities. We aim to provide a system for selecting a collaborator for joint research or an expert on a given list of topics. We have improved a recommender system for finding possible collaborator with respect to research interests and predicting quality and quantity of the anticipated publications. Our system is based on a co-authorship network derived from the bibliographic database, as well as content information on research papers obtained from SJR Scimago, staff information and the other features from the open data of researchers profiles. We formulate the recommendation problem as a weighted link prediction within the co-authorship network and evaluate its prediction for strong and weak ties in collaborative communities.

**Keywords:** Recommender systems · Co-authorship network
Scientific collaboration

## 1 Introduction

In a modern scientific community it is important to know the trends that provide the significant impact on the research fields. However, it is not easy to read hundreds of papers to become familiar with the new topics and small improvements in many related fields of study. The most natural way to select relevant and most valuable articles is by ordering a list of articles obtained by keywords query from some bibliography database according to a citation index or other centrality metrics for measuring simultaneously influence of the author and the paper on the respected research area [1]. In fact, such a method does not take into account the author professional skills, his respective research community and ability to publish his research at the international level. One of the first methods for selecting analyzing research community were made by Newman in [2,3], where the author ordered authors according to the collaboration and centrality metrics in the co-authorship network.

W. M. P. van der Aalst et al. (Eds.): AIST 2017, LNCS 10716, pp. 404–410, 2018.
https://doi.org/10.1007/978-3-319-73013-4_37

Clustering approach for a co-authorship network of researchers who studied a particular disease was presented in [4]. The authors of [5] gave a representation of finance network analysis using similar methods. In [6,7], the authors studied dependencies between citation indexes (predicted citations in [8]) and centralities in a co-authorship network. Data mining approaches for extracting significant features from the co-authorship networks specified to different research areas were presented in [2,9]. Overall evaluation of methods and applications of network analysis were described in [10].

In this paper, we study a co-authorship network based on co-authorship network while one or more among the coauthors belong to the National Research University Higher School of Economics (HSE). The core idea of the research is to apply network analysis [10] and topic modelling [11] of research papers for the problem of extracting research interests of HSE workers and their respective skills based on the qualitative and quantitative data of their publications. Such obtained system then could be applied for automatising of expert search [12], building recommender system for searching a collaborator or scientific adviser [13], and simple search engine who could possess knowledge and skills related to a given description.

In what follows, we describe in details the process of evaluating work-in-progress recommender system based on co-authorship network and information on the staff units profiles of each author from HSE.

## 2 Data Processing and Problem Formalisation

As a starting point we took the database of all the records from the NRU HSE publication portal [14]. Here, we describe the process of cleaning original database:

1. All the duplicate records were merged under assumption that any conflict could be resolved by choosing the data from verified record at the portal.
2. All the missing fields were omitted during computational part of filed with median over respective category of articles and authors.
3. All the conflicts of different author First/Last Name representation were solved under simple logistic regression model based on the number of common co-authors in a given dataset.

After cleaning stage, we build five layers of the co-authorship network with authors as network actors and edges connecting authors with $k$ jointly published research articles, $k = 1, \dots, 5$. The number 5 was chosen as the maximal number under which the network does not degenerate to the large number of small connected components of average sizes 1–2.

Next, we import features related to both, actors and ties between them. For each tie in the network representing co-authored publication, we added all the attributes of the publication the university portal [14] and imported subject areas and categories from Science Journal Ranking [15,16]. We measure publication quality as its respective quartile in SJR ranking for the publication year, computed as maximal (or average) over different categories per journal. In the future work we aim to include quartiles from the Web of Science Core Collection and provide unified algorithm of measuring relevant quartile in accordance to NRU HSE Science Fund Policy. One of the key features of new system will be choosing relevant quartile with respect to subject category choice over several research areas provided for indexed journal. Information about network actors, such as administrative unit position, declared author research interests and additional interests derived from topic modelling BigARTM system [17] over author's research publications were also included as node features.

We use the co-authorship network for various research problems related to collaboration patterns: dynamics of the number of papers as a personal progress indicator; dynamics of the number of collaborators as a representation of qualification and networking of the researcher; dynamics of the network communities density as an attribute to measure of team-working; impact of the research area and administrative units on collaborative and numerical publication patterns; using text mining, find a relevant expert or a collaborator based on his/her research interests and measured quality of co-author collaboration in international publications.

The recommender system gives a list of suggested candidates ordered by the inner metric of successful collaboration. More detailed, for a selected author the system generates a ranked list of authors whose papers could be relevant to him, and authors themselves could be good candidates for collaboration.

## 3 Measuring Similarity of Interests and Authors

We consider the problem of finding authors with similar interests to a selected one. In terms of network analysis, we study the problem of recommending similar author as a link prediction and use similarity between authors as model features. We choose well-known similarity scores described in [18].

We use *Common neighbors*, *Jaccard's coefficient*, *Adamic/Adar*, *Graph distance* similarity scores as baseline. In order to define similarity of actors by their known features from HSE staff information and publication activity represented by centralities of co-authorship network, we define additional content-based and graph-based features.

Nodes of the graph correspond to authors and, hence, have binary attributes, relation with NRU HSE and whether has administrative position, staff position, full-time status, and qualitative attributes, first name/last name, department hierarchy, and centrality metrics from the co-authorship network as metrics of influence. Edge attributes includes title and data of publication, journal quartile, weight, subject area and category, whether it was indexed by Scopus.

We computed cosine similarity for a vector consisting of normalized values of the feature parameters and "interests" metric as a normalized number of common journal SJR subject areas and categories for the journals, in which the original research article were published.

## 4 Training Model

In order to create proper training set for the person-based recommender system, we first need to select the list of potential candidates with similar interests.

We start with previous co-authors and staff members from the same HSE administrative unit. We construct feature vectors for their respective staff units by computing descriptive statistics based on research activity criteria and simple summarization of their respective researchers profiles related to different time intervals and qualities of publications.

We add to the list of candidates those administrative units with the difference between feature vectors less than the median of the distances between all the pairs of staff unit feature vectors representations, defined by the following features: the number of authors and papers, the ratio of the number of papers (from the department) by the ratio of the authors (from the department), the number of articles indexed in Scopus (over 3 last years) (published in Q1 and Q2 journal quartiles) (divided by the number of authors) (from the department), the average number of co-authors, the number of connected components, size of the greatest connected component (GCC), average distance, graph diameter, radius and density, average local clustering coefficient, the number of lecturers/senior lecturers/assistant professors/professors with their respective numbers of papers published (over the last 3 years), average rating of senior lecturers/assistant professors/professors from 0 to 30. The rating was calculated by the formula $\min(15 \cdot N, 30)/\min(15 \cdot N, 30)/\min(10 \cdot N, 30)/\min(6 \cdot N, 30)$, where $N$ is the number of publications over the last 3 years. Each of the parameters was chosen according to either publishing activity criteria specified for different staff categories, or correlation between subgraph publishing patterns and their descriptive statistics.

In what follows, we use topic modelling for initial candidate research papers and find all the similar authors in HSE co-authorship network based on cosine metric of their research interests with the taxonomy obtained from hierarchical clustering of interests in respective co-occurrence network.

In order to catch structure of the network we used five methods of community detection on the co-authorship network: label propagation, fastgreedy, louvain, walktrap, infomap [19], and added candidates from the obtained clusters, to which original person belonged (Table 1).

The number of communities appears to be quite stable. Finally, we removed all non-HSE authors due to lack of information on their activity and status outside of HSE collaboration.

## 5 Recommender System

We used logistic regression with lasso regularization on normalized feature vectors to predict new links [20]. The parameter for regularization was chosen via model fitting by maximising accuracy of the model [21].

**Table 1.** The number of clusters obtained by different algorithm

| Weight | >0 | >1 | >2 | >3 | >4 | >5 |
|---|---|---|---|---|---|---|
| Label | 2265 | 1313 | 937 | 735 | 582 | 453 |
| Fastgreedy | 1120 | 954 | 748 | 608 | 511 | 406 |
| Louvain | 1102 | 943 | 744 | 605 | 509 | 406 |
| Walktrap | 2233 | 1270 | 890 | 673 | 542 | 428 |
| Infomap | 1988 | 1215 | 876 | 684 | 554 | 432 |

For a given researcher, we form subgraph of candidates from the previous section as a training set with the edges induced by original co-authorship network. We construct logistic regression model for each of the groups, taking as positive examples links in the chosen subgraph, and the same number of negative examples as missing links in order to keep the balance in the model, similar to [22].

We train our model on the "strong" co-authorship networks with each edge appearing only if up to $k = 2$ to $k = 5$ papers were written together and obtained a series of subgraphs to be used in evaluation (see [13]). For all the pairs of "weak" and "strong" subgraphs we prepare test set as links from the difference of these graphs and the same number of missing links from the links difference with features taken from the stronger subgraph. We calculated average error rates for test and train sets over all pairs of thresholds values of $k$ (see Table 2). The area under the rock curve (AUC) and F1-measure are high, therefore, normalized lasso logistic regression was sufficient for binary classification.

**Table 2.** Similarity metrics

| | Precision | Recall | Accuracy | F1-measure | AUC |
|---|---|---|---|---|---|
| Train data | 0,93 | 0,99 | 0,96 | 0,96 | 0,99 |
| Test data | 0,91 | 0,87 | 0,91 | 0,90 | 0,94 |

## 6 Conclusion

We improved a recommender system [13] providing ranked list of candidates for collaboration based on HSE co-authorship network and database of publications. The recommender system demonstrates promising results on predicting new collaborations between existing authors and the accuracy of the system was improved by adding topic modelling component for extracting research interests from the original papers. The recommendations could also be made for a new author, who should state research interests and/or load his research papers for topics extraction.

We are looking forward to the evaluation of our system for several tasks inside the NRU HSE (though, it could be applied to any other research community), such as:

- finding an expert based on text for evaluation;
- matchmaking for co-authored research papers with novice researchers;
- searching for scientific adviser based on co-authorship network and the probability of publication in co-authorship with a student;
- searching for collaborators on specific grant proposal.

An application of this system may help stating the University policy to support novice researchers and increase their publishing activity or even, estimate collaboration between the University staff units.

**Acknowledgements.** The work was supported by the Russian Science Foundation under grant 17-11-01294 and performed at National Research University Higher School of Economics, Russia.

## References

1. Liang, Y., Li, Q., Qian, T.: Finding relevant papers based on citation relations. In: Wang, H., Li, S., Oyama, S., Hu, X., Qian, T. (eds.) WAIM 2011. LNCS, vol. 6897, pp. 403–414. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23535-1_35
2. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proc. NAS **101**(suppl 1), 5200–5205 (2004)
3. Newman, M.: Who is the best connected scientist? A study of scientific coauthorship networks. Complex Netw. **650**, 337–370 (2004)
4. Morel, C.M., Serruya, S.J., Penna, G.O., Guimarães, R.: Co-authorship network analysis: a powerful tool for strategic planning of research, development and capacity building programs on neglected diseases. PLoS Negl. Trop. Dis. **3**(8), e501 (2009)
5. Cetorelli, N., Peristiani, S.: Prestigious stock exchanges: a network analysis of international financial centers. J. Bank. Finance **37**(5), 1543–1551 (2013)
6. Li, E.Y., Liao, C.H., Yen, H.R.: Co-authorship networks and research impact: a social capital perspective. Res. Policy **42**(9), 1515–1530 (2013)
7. Yan, E., Ding, Y.: Applying centrality measures to impact analysis: a coauthorship network analysis. J. IST Assoc. **60**(10), 2107–2118 (2009)
8. Sarigöl, E., et al.: Predicting scientific success based on coauthorship networks. EPJ Data Sci. **3**(1), 9 (2014)
9. Velden, T., Lagoze, C.: Patterns of collaboration in co-authorship networks in chemistry-mesoscopic analysis and interpretation. In: ISSI 2009 (2009)
10. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)
11. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 448–456. ACM (2011)
12. Mimno, D., McCallum, A.: Expertise modeling for matching papers with reviewers. In: Proceedings of the 13th ACM SIGKDD IC, pp. 500–509 (2007)
13. Makarov, I., Bulanov, O., Zhukov, L.: Co-author recommender system. In: Kalyagin, V., Nikolaev, A., Pardalos, P., Prokopyev, O. (eds.) Springer Proceedings in Mathematics and Statistic, vol. 197, pp. 1–6. Springer, Cham (2017)

14. Powered by HSE Portal: Publications of HSE (2017). http://publications.hse.ru/en. Accessed 9 May 2017
15. González-Pereira, B., Guerrero-Bote, V.P., Moya-Anegón, F.: A new approach to the metric of journals' scientific prestige: the SJR indicator. J. Informetr. **4**(3), 379–391 (2010)
16. Guerrero-Bote, V.P., Moya-Anegón, F.: A further step forward in measuring journals' scientific prestige: the SJR2 indicator. J. Informetr. **6**(4), 674–688 (2012)
17. BigARTM contributors: BigARTM v0.8.2, December 2016. https://doi.org/10.5281/zenodo.288960
18. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. IST Assoc. **58**(7), 1019–1031 (2007)
19. Lancichinetti, A., Fortunato, S.: Community detection algorithms: a comparative analysis. Phys. Rev. E **80**(5), 056117 (2009)
20. Meier, L., Van De Geer, S., Bühlmann, P.: The group LASSO for logistic regression. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **70**(1), 53–71 (2008)
21. Wainwright, M.J., Ravikumar, P., Lafferty, J.D.: High-dimensional graphical model selection using $l_1$-regularized logistic regression. Adv. Neural Inf. Process. Syst. **19**, 1465 (2007)
22. Beel, J., et al.: Research paper recommender system evaluation: a quantitative literature survey. In: Proceedings of the International Workshop on RepSys 2013, pp. 15–22. ACM, New York (2013)