# Co-author Recommender System

Ilya Makarov and Oleg Bulanov and Leonid E. Zhukov

**Abstract** Modern bibliographic databases contain significant amount of information on publication activities of research communities. Researchers regularly encounter challenging task of selecting a co-author for joint research publication or searching for authors, whose papers are worth reading. We propose a new recommender system for finding possible collaborator with respect to research interests. The recommendation problem is formulated as a link prediction within the co-authorship network. The network is derived from the bibliographic database and enriched by the information on research papers obtained from Scopus and other publication ranking systems.

## 1 Introduction

Nowadays, researchers have to deal with hundreds of papers to become familiar with their fields of study. However, the number of the papers exceeds human abilities to read them all. The most common way to select relevant articles is by sorting a list of all articles according to a citation index and choosing some articles from the top of the list. However, such a method does not take into account the author professional specialization. Another way of selecting suitable articles is to choose articles of well-known authors. A more advanced methods was proposed by Newman in [13, 14], where he ranked authors according to the collaboration weight or values of centrality metrics such as degree and betweenness in the co-authorship network. In [12] authors decided to cluster authors at a co-authorship network who studied a particular disease, while in [4] authors gave a representation of finance net-

National Research University Higher School of Economics, 125319, Kochnovskiy Proezd 3, Moscow, Russia
· Ilya Makarov e-mail: iamakarov@hse.ru,revan1986@mail.ru
· Oleg Bulanov e-mail: oleggl500@gmail.com
· Leonid E. Zhukov e-mail: lzhukov@hse.ru

work analysis. In  [9, 17] the authors studied correlation between citation indexes and centrality measures in a co-authorship network and in [15] predicted citation indexes from centrality metrics. There are also exist numeral publication studying general features of co-authorship networks in various science fields [13, 16, 18]. The methods and applications of network analysis were described in [19].

In this paper we present a co-authorship network based on papers co-authored by researchers from the National Research University Higher School of Economics (HSE). HSE authors often have publications in collaboration with non-HSE authors, so it is necessary to include such authors to the network but the recommender system gives recommendations only among HSE researchers. Non-HSE authors were added in order to calculate network nodes metrics more precisely.

We started by taking a relational database of all publications written by NRU HSE authors. We cleaned the database by removing duplicate records and inconsistencies, unified author identifiers and filled in missing data. An undirected co-authorship graph was constructed with authors as graph nodes and edges containing lists of jointly published papers. We added all publication attributes from the university portal and subject areas and categories from Scopus Journal Ranking [5, 6]. Publication quality was taken from its quartile in SJR ranking for the publication year, computed as maximal (or average) over different categories per journal. Information about authors administrative units and declared author research interests was also included as node features.

The co-authorship graph can help to answer a variety of research questions about collaboration patterns: distribution of number of papers written by an author and distribution of number of collaborators, density of graph communities, dependency on research area and administrative units, evolution of collaborations over time etc. We use the graph to power a co-author recommender system. The system gives recommendations of authors that have interests similar to the chosen author or whose co-authorship pattern is similar to that of the author. More specifically, 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.

## 2 Author Similarity Score

Let us consider  problem of finding authors with similar interests to a selected author. We formulate this recommendation problem as a problem of link prediction in the network and use similarity between network nodes for prediction. The comparative analysis of network similarity metrics is provided in [10].

All similarity metrics can be divided into two types. The first four metrics from the Table 1 are standard similarity metrics described in [10]. Since nodes of the co-authorship network represents known authors from HSE, one can also define additional content based features for similarity metrics between authors'. We used the following content based features: ithe number of fields of journals, the number of papers, the number of papers in journals of high quartiles, the number of papers during past 3 years, etc., local clustering coefficients, degree, betweenness and eigenvector centrality metrics, position and seniority. We calculated cosine similarity for a vector consisting of normalized values of the feature parameters and “interests” metric for the journals where papers where published.

**Table 1** Similarity metrics

| Similarity metric | Definition |
|---|---|
| common neighbors | $sim(v_i, v_j) = \|N(v_i) \cap N(v_j)\|$ where $N(v)$ is number of neighbours and $v$, $v_i, v_j$ are nodes |
| Jaccards coefficient | $sim(v_i, v_j) = \frac{\|N(v_i) \cap . N(v_j)\|}{\|N(v_i) \cup N(v_j)\|}$ |
| Adamic/Adar | $sim(v_i, v_j) = \sum_{v \in N(v_i) \cap N(v_j)} \frac{1}{\ln \|N(v)\|}$ |
| graph distance | length of the shortest path between $v_i$ and $v_j$ |
| cosine | $sim(a, b) = \frac{(a,b)}{\|\|a\|\| \cdot \|\|b\|\|}$ |
| interests | normalized number of common journal SJR areas |

## 3 Choosing subgraphs for a training set

HSE co-authorship network contains nodes that represent authors from different fields of study and departments. In [13] Newman showed that researchers from different scientific areas form new connections differently. So we first created overlapping groups of authors from similar affiliation and scientific interests.

We start with forming department subgraph, defined by unit staff membership for HSE co-authors. We then construct a feature vector for a department consisting of over 30 different descriptive statistics of the department subgraph, quantities of publications and normalized publications activity of the researchers with respect to different time intervals and quartiles. We considered two departments *similar*, if the norm of the difference between their feature vectors is less than the median of the distances between the pairs of the feature vectors for all departments.

We select candidates from each department by the following procedure. Initially, all the authors from the department, as well as the authors from similar departments are considered as candidates. Every author with the same areas of journals as those of the selected author’s department is also added as a candidate. We used five methods of community detection on the co-authorship network, such as label propagation [7], fastgreedy [11], louvain [3], walktrap [8], infomap [2], and created five candidate sets by unifying the previous set with all the found communities containing authors from the previous set. Finally, all non-HSE authors were removed from these sets. The Euler diagram of the obtained groups is shown at the Figure 1.

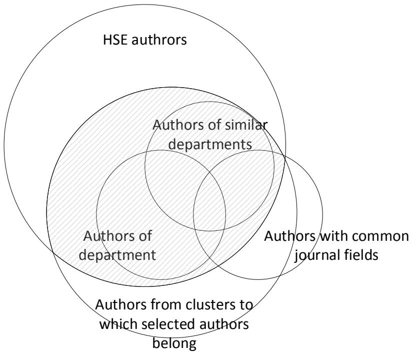
HSE authrors
Authors of similar departments
Authors of department
Authors with common journal fields
Authors from clusters to which selected authors belong

**Fig. 1** Euler diagram of group of authors among which recommendations will be given

## 4 Recommender system

We used linear regression on normalized feature vectors to predict new links. We applied lasso regularization and choose high regularization parameter. Equation 1 shows a lasso regression, where $X$ is matrix of similarity metrics values, $Y$ is a vector indicating links' presences, $\lambda$ is a regularization parameter and $N$ is a number of similarity metrics.

$$\begin{cases} \frac{1}{N}(Y - X\theta)^T(Y - X\theta) + \lambda||\theta||_1 \longrightarrow \min_\theta \\ ||\theta|| < \lambda \end{cases} \tag{1}$$

Let us describe the process of constructing the recommender system similar to [1]. For a given researcher we form corresponding subgraphs for training sets from the previous section and choose only that contained our researcher. We construct linear 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 the same subgraph. For a fixed group we choose one community detection method with the highest precision among five corresponding to different clustering methods (see Algorithm 1).

```
Algorithm 1: Algorithm of constructing a recommender system
Data: $N$ - co-authrship network
Result: $\{\theta\}$ - regression coefficients for each group, $\{G\}$ - groups
begin
    $\{G\} \longleftarrow$ all groups, five for each department
    for $G \in \{G\}$ do
        $X_{train} \longleftarrow$ features of links and the same number $n$ of features of nonexistence links
        $Y_{train} \longleftarrow (1\dots1, 0\dots0)$ with $n$ units and $n$ zeros
        $(\theta, \lambda) \longleftarrow$ calculate regression coefficients fitting $\lambda$ for highest precision
    $\{G\} \longleftarrow$ select one group for each department
```

After we compute all linear models for each of the groups, we can describe the scheme of making a recommendation. At first, we choose a group with the highest precision corresponding to one of the departments, which the selected author belongs to. A group should be fixed because an author may belong to several departments simultaneously. Secondly, we take a normalized vector of predictions from linear regression. We provide a ranked list of recommendation ordered by the predicted values that are greater than 0.5.

## 5 Results

We predicted links between those authors that have written from $k = 1$ to $k = 5$ papers together and obtained a series of, so called, "strong" subgraphs to use in cross-validation. For all the pairs of $k_1/k_2$-subgraphs we calculated the predictions for the "stronger" subgraph. For each group, we build two subgraphs induced by authors from a group in the corresponding stronger subgraph. We find the difference of the link sets for $k_1$ and $k_2$ stronger subgraphs ($k_1 < k_2$). If the difference is not empty, we prepare the test sample as a set of links from the difference and the same number of missing links from the links difference with features taken from the stronger subgraph, otherwise, we change a group. For all the groups we calculate 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 regression is sufficient for binary classification.

**Table 2** Similarity metrics

| | Precision | Recall | Accuracy | F1-measure | AUC |
|---|---|---|---|---|---|
| train data | 0,916251 | 0,991259 | 0,9467785 | 0,949979 | 0,990913 |
| test data | 0,901435 | 0,867798 | 0,8733743 | 0,870438 | 0,923516 |

## 6 Conclusion

We developed a recommender system based on HSE co-authorship network. The recommender system demonstrates promising results on predicting new collaborations between existing authors and can fasten the process of finding collaborators and relevant research papers. The the recommendation system can be also used for new authors, who do not have any connections to HSE community. A further analysis of the co-authorship network may help stating university policy to support novice researchers and increase their publishing activity and estimate collaboration between the university departments. Though tested on HSE co-authorship network, the approach can be easily applied to other networks.

## References

1. Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., Nurnberger, A.: Research paper recommender system evaluation: a quantitative literature survey, ACM RepSys, 15-22, (2013)
2. Bergstrom, C.T., Rosvall, M.: Maps of random walks on complex networks reveal community structure, PNAS, vol. 105 (4), 1118–1123 (2008)
3. Blondel, V.D., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech., vol. 2008 (10), 12p. (2008)
4. Cetorelli, N., Peristiani, S. Prestigious stock exchanges: A network analysis of international financial centers, Journal of Banking & Finance, vol. 37 (5), 1543–1551 (2013)
5. Gonzlez-Pereira, B., Guerrero-Bote, V.P., Moya-Anegn F.: A new approach to the metric of journals scientific prestige: The SJR indicator, J. of Informetrics, vol. 4 (3), 379–391 (2010)
6. Guerrero-Bote, V.P., Moya-Anegn F.: A further step forward in measuring journals scientific prestige: The SJR2 indicator, Journal of Informetrics, vol. 6 (4), 674–688 (2012)
7. Kumara, S., Raghavan, U. N., Albert, R.: Near linear time algorithm to detect community structures in large-scale networks, Phys Rev E., vol. 76 (3), 12p. (2007)
8. Latapy, M., Pons, P.: Computing communities in large networks using random walks, Computer and Information Sciences - ISCIS 2005, LNCS, vol. 3733, 284–293. (2005)
9. Li, E.Y., Liaoa, C.H., Yenb, H.R.: Co-authorship networks and research impact: A social capital perspective, Research Policy, vol. 42 (9), 1515-1530 (2013)
10. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks, J. of the American Society for Information Science and Technology, volume 58 (7), 1019–1031 (2007)
11. Moore, C., Clauset, A., Newman, M. E. J.: Finding community structure in very large networks, Phys. Rev. E, vol. 70, 6p.(2004)
12. Morel, K.M., Serruya, S.J., Penna, G.O., Guimaraes, R.: Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases, PLOS neglected tropical diseases, vol. 3 (8), 1–7 (2009)
13. Newman, M. E. J.: Coauthorship networks and patterns of scientific collaboration, PNAS, vol. 101 (suppl 1), 5200-05205 (2004)
14. Newman, M. E. J.: Who is the best connected scientist? A study of scientific coauthorship networks, Complex Networks, LNPh, vol. 650, 337–370 (2000)
15. Sarigl, E., Pfitzner, R., Scholtes, I., Garas, A., Schweitzer, F.: Predicting Scientific Success Based on Coauthorship Networks, EPJ Data Science, vol. 3 (1), 9p. (2014)
16. Velden, T., Lagoze, C.: Patterns of Collaboration in Co-authorship Networks in Chemistry - Mesoscopic Analysis and Interpretation, ISI - 2009, vol. 2, 12p. (2009)
17. Yan, E., Ding, Y.: Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis, J. Am. Soc. Inf. Sci. Technol., vol. 60 (10), 2107-2118 (2009)
18. Zervas, P., Tsitmidell, A., Sampson, D.G., Chen Nian-Shing, Kinshuk: Studying Research Collaboration Patterns via Co-authorship Analysis in the Field of TeL: The Case of ETS Journal, Journal. Educational Technology and Society, vol. 17 (4), 1–16 (2014)
19. Wasserman, S., Faust, F. Social Network Analysis. Methods and Applications, Cambridge University Press, 825p. (1994)