Recommending Co-authorship via Network Embeddings and Feature Engineering

The case of National Research University Higher School of Economics

Ilya Makarov
National Research University Higher School of Economics
International Laboratory for Applied Network Research
Moscow, Russia
iamakarov@hse.ru

Olga Gerasimova
National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence
Moscow, Russia
olga.g3993@gmail.com

Leonid E. Zhukov
National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence
Moscow, Russia
leonid.e.zhukov@gmail.com

ABSTRACT

Co-authorship networks contain hidden structural patterns of research collaboration. While some people may argue that the process of writing joint paper depends on mutual friendship, research interests and university policy, we proved that given a temporal co-authorship network one could predict the quality and quantity of future research publications. We compare existing graph embedding and feature engineering methods, presenting combined approach for constructing co-author recommender system formulated as link prediction problem. We evaluate our research on a single university publication dataset providing meaningful interpretation of the obtained results.

CCS CONCEPTS

• **Information systems → Collaborative search**; **Link and co-citation analysis**; *Recommender systems*; • **Human-centered computing → Collaborative and social computing design and evaluation methods**; *Collaborative interaction*; *Social recommendation*; *Social networks*; Empirical studies in collaborative and social computing;

KEYWORDS

Co-authorship Networks, Recommender Systems, Machine Learning, Graph Embeddings, Link Prediction

1 INTRODUCTION

Modern research community has been overflowed by a large amount of international conferences and journal articles with constantly increasing number of relevant research areas and papers. It is important to know the trends in respected research fields, while not reading hundreds of papers to become familiar with the new topics and small improvements in many related fields of study. The simplest way to select most valuable articles is by ordering a list of articles obtained by keywords query from some bibliography database, taking into account citation index or other centrality metrics for measuring influence of the author and the paper on the respected research area [33]. In fact, such an approach does not include the information on the author professional skills, his/her research community and ability to publish results of the research at high-impact journals. One of the first methods for analyzing research communities was suggested by Newman in [47], [46], where the authors were ordered according to the collaboration and centrality metrics in the co-authorship network.

Unsupervised learning for cluster structure of researchers' co-authorship network who studied a particular disease was presented in [45]. Another study of finance network with similar methods was suggested in [9]. In [30, 71], the authors studied dependencies between citation indexes, based on predicted citations [54], and centralities in a co-authorship network. Data mining approach for feature engineering specified to different research areas was presented in [47, 64]. Overall evaluation of applied network analysis methods was described in [68].

Link prediction is a problem of predicting links in temporal networks and missing link data in complex networks. The algorithms can be used to extract missing link information, identify abnormal interaction activity, model network evolving processes and estimate the most probable persons to be connected with.

We formulate the problem of recommending co-author as a link prediction problem [31], in which the model should be able to predict whether a pair of nodes in a network would have an edge connecting them and what will be the parameters of such an edge in a given time period. Link prediction is used in many applied problems, such as web hyper-link prediction [1, 77], social dating [17], genomics [19, 63], real-world friend search in social networks [3], and digital libraries [15, 41]. A survey on link prediction applications was published in [57]

Recently, the major advancement of vectorized natural language text representations reaches the network science area, suggesting new approaches in representation learning of graph embeddings [49]. The consequent improvement of network embedding models [10, 24, 49, 59] leaded to state-of-art performance on clustering, multi-class node classification and link prediction tasks. However, existing methods prioritize gaining invariant under structural equivalence and homophily network properties, but skip the information

JCDL'18, June 2018, Fort Worth, Texas, USA

ACM ISBN .
https://doi.org/

outside of the network, which may contain important information, such as the affiliated texts, images, labels and different social network user attributes.

In this paper, we study a co-authorship recommender system based on co-authorship network while one or more among the coauthors belong to the National Research University Higher School of Economics (HSE). We use machine learning techniques to predict new edges in temporal structure of co-authorship network, while extracting author and network features, providing different word [29, 65] and graph [24, 70] embeddings for network actors, their research papers and attributes. We compare our approach with state-of-art algorithms for link prediction problem and show that combined feature engineering and network emebedding outperform each approach separately.

Such obtained system could be applied for expert search [44] ranked according to research influence of co-authorship network actors, building recommender system for searching a collaborator or scientific adviser [39, 40], and building a simple search engine of relevant research publications.

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

2 RELATED WORK

2.1 Link Prediction

The link prediction problem was first formulated in [35], where the approach based on measures for analyzing the nodes proximity in network was suggested. The provided experiments on large co-authorship networks showed that information on future links can be extracted from network structure alone.

Supervised approach was proposed in [2]. Unsupervised feature selection was suggested in [58]. The authors of [56] suggested relational learning algorithm using matrix factorization. In [22], a temporal link prediction involving node proximity and node content was made also using matrix factorization technique. The feature learning based on latent learning was presented in [75]. In [32], a Conditional Temporal Restricted Boltzmann Machine was used for modelling the dynamics of network links. In [76], a multi-network link identifier was proposed to describe the multi-network link formation.

In [26], the authors presented a survey on link prediction methods categorized on three types of the models: feature extraction for binary classification, Bayesian graphical models, and an approach of dimensionality reduction from linear algebra matrices operations. The survey on classical link prediction was made in [38].

In what follows, we describe two most studied approach for link prediction task for large-scale networks, providing feasible solutions based on state-of-art techniques for Homophily and Graph Embedding, while leaving the up-to-date survey on link prediction to a reader [67].

2.2 Actor Homophily and Graph-based Recommender Systems

The simplest baseline solution is based on Common Neighbors or other network similarity scores [36]. A number of coded solutions could be found as a standard part of NetworkX Python package [13]. In [21], the authors showed that the similarity measures are highly influenced by the type of network, and could provide both, good and bad results depending on network properties.

The impact of the attribute-based formation of social networks rather than only self-organizing (according to power-law and preferential attachment) was considered in [52]. The main idea to consider attribute properties is in the impact of attribute homophily, which is one of the most known empirical observation from the real world [42]. The different assortative mixing patterns based on age, sex, group interest, and even romance relations, have been studied as a part of emerging social network analysis [68]. Mostly, all these approaches require manually engineered feature selection from the problem domain. A modern approach for computationally efficient link prediction via feature engineering was suggested in [18].

A series of works on recommender systems via link prediction problem [11, 31, 37] started the theory of graph recommender showed state-of-art performance based on extracting structure information from the networks.

In [28], the authors studied the effect of homophily in a university community, considering temporal co-authorship network accompanied with staff and affiliation information. The authors concluded that not only structural proximity, but rather orientation to form links with actors possessing similar attributes highly impacts the social network structure over time.

We follow this idea while studying the large HSE research community creating staff units and research interests attributes as a foundation for actor-based homophily.

2.3 Graph Embeddings

In supervised machine learning problems, one has to construct a feature vector in order to represent network nodes and edges. Usually, hand-engineering features based on domain expert knowledge would require. Their applicability would be highly influenced by particular tasks, as well as by the quality of an expert and size of the data. Nowadays, the development of digital technologies leads to aggregating big data that could not be manually engineered, that is why the theory of finding hidden representations (see [5]) has impacted the whole field of machine learning and artificial intelligence community. The challenge in feature learning via optimization problem is consisted in defining loss function, while simultaneously supporting robustness of inner representation and domain generalization. Local Linear Embedding [53], IsoMAP [61], Laplacian Eigenmap [4], Spectral Clustering [60], MFA [72] and GraRep [7] were the first attempts to embed the graphs into vector space based on several locality measures. However, development of representation learning for networks has major drawbacks when considering dimensionality reduction techniques such as PCA [20], SVD [14], MDS [62], LDA [6], etc., involving eigenvalue-based decomposition of the corresponding network matrix, which is not computationally efficient for large-scale networks and have bad results in many classification tasks failing to preserve robustness of the model for noised edge data of the network.

Recent papers on network embeddings concentrate mostly on improving performance on several typical machine learning tasks

based on describing network in terms of random walking. In DeepWalk [49] and node2vec [24] algorithms, the authors use Skip-gram [43] based model to simulate breadth-first sampling (BFS) and depth-first sampling (DFS). More weak representations based on the first-order and second-order nodes proximity were suggested in LINE [59] and SDNE [66] models.

In [8], research group make more thorough investigation of node2vec [24] graph embedding finding certain drawbacks in the suggested fine-tuning of hyper-parameters for balancing BFS and DFS. In [70], the authors argue that node2vec model should be accompanied with vectorized model of the whole network.

While the connections between node and edge embedding are still to be studied (as mentioned in [24]), we concentrate on applying graph and feature embeddings methods to a link prediction task on a particular network of HSE researchers co-authorship network with many additional attributes on actors. Purely network-based methods fail to include the information from actors obtained from the other sources resulting in decrease of efficiency of network embeddings. In what follows, we consider several approaches aimed to combine structure and attribute information of network actors.

2.4 Combined Deep Learning Architectures

Several approaches were suggested recently to incorporate actor's label and text information, such as text-associated DeepWalk model TADW [73] and more advanced label informed attributed network embedding LANE [27]. In tri-party deep network representation TriDNR [48], the authors proposed separately learn embeddings from DeepWalk [49] and label and content embedding modeled by Doc2Vec [29] model. On the contrary social network embedding framework SNE [34] learns representations for social actors by preserving both the structural equivalence and node attribute proximity simultaneously in end-to-end neural network architecture. The semi-supervised learning for network embedding has been studied in [69, 74].

3 DATA PROCESSING AND FEATURE GENERATION

In this section we present information on data extraction from digital archive of HSE publications and feature selection process according to author homophily features and structural graph emebeddings.

3.1 HSE Publication Digital Library

As a starting point, we took the database of all the records from the NRU HSE publication portal [50]. The data source that we use is the electronic library of scientific papers published in co-authorship with HSE researchers and placed on this site by one of the co-authors.

The database contains infromation on 6996 HSE authors published 30668 research papers, including journals, conference reports, proceedings, monographs, books, preprints and scientific reviews. The portal contains tools for uploading data for a fixed time period. The fields of the database contain information of a title, a list of authors, keywords, abstract, data, place, journal/conference and publishing agency and several attributes of the paper and journal issue, as well as information on indexing in Scopus, Web of Science (WoS) Core Collection and Russian Science Citation Index (RSCI).

Unfortunately, the database has not been integrated with any research paper digital libraries and does not support automatically uploading of bibliography descriptions, such as *.ris, *.enw or *.bib files, which implies in noisy data representation with heterogeneous descriptions, abbreviated author homonimy and incomplete information on the publications. The full texts are usually unavailable under security properties chosen by the corresponding authors, while the rear paper abstracts are not sufficient to be used for topic modelling. In such a case, we include the information on research interests from the Scopus subject categories of the journals, in which the author has published research articles, and manually input at the personal web-page of researcher research interest list according to RSCI categorization.

In what follows, we briefly describe the process of cleaning the 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.

A part of this network representing the ability to generalize corporate relations by visualizing co-authorship subgraph of the most publishable person (rector of HSE, in fact) and his co-authors (representing leading research laboratories and administrative units) could be seen at Figure **??**.

3.2 Author Disambiguation

World-wide, many researchers with the same name or initials work in the same organization, that is why we need to resolve the conflicts of potential author name multiple matching to the existing workers entries, which is called disambiguation. In certain cases when the author is mentioned as HSE worker the database contain the link to the personal web page, which is unique to all the staff members. A number of ambiguous author descriptions was approximately 2% of the whole database.

The authors, which have published their work in English and Russian languages, have been transliterated into English equivalent and matched under assumption of small Levenshtein distance measure for Last name and initials.

The disambiguation of the authors with the same initials was solved using simple logistic regression model based on the number of common co-authors in a given dataset, similarity score on the vectors of research interests from both candidates and the category of the published article with ambiguous author description.

Finally, we removed all non-HSE authors due to lack of information on their activity and status outside of HSE collaboration.

3.3 Measuring Publication Quality

After cleaning stage, we import features related to both, actors and links between them. We have calculated a descriptive statistics of HSE research publications showing that despite the strategic direction to appear in Top-100 world universities, the major department structures lack the high-impact papers, and research activity of

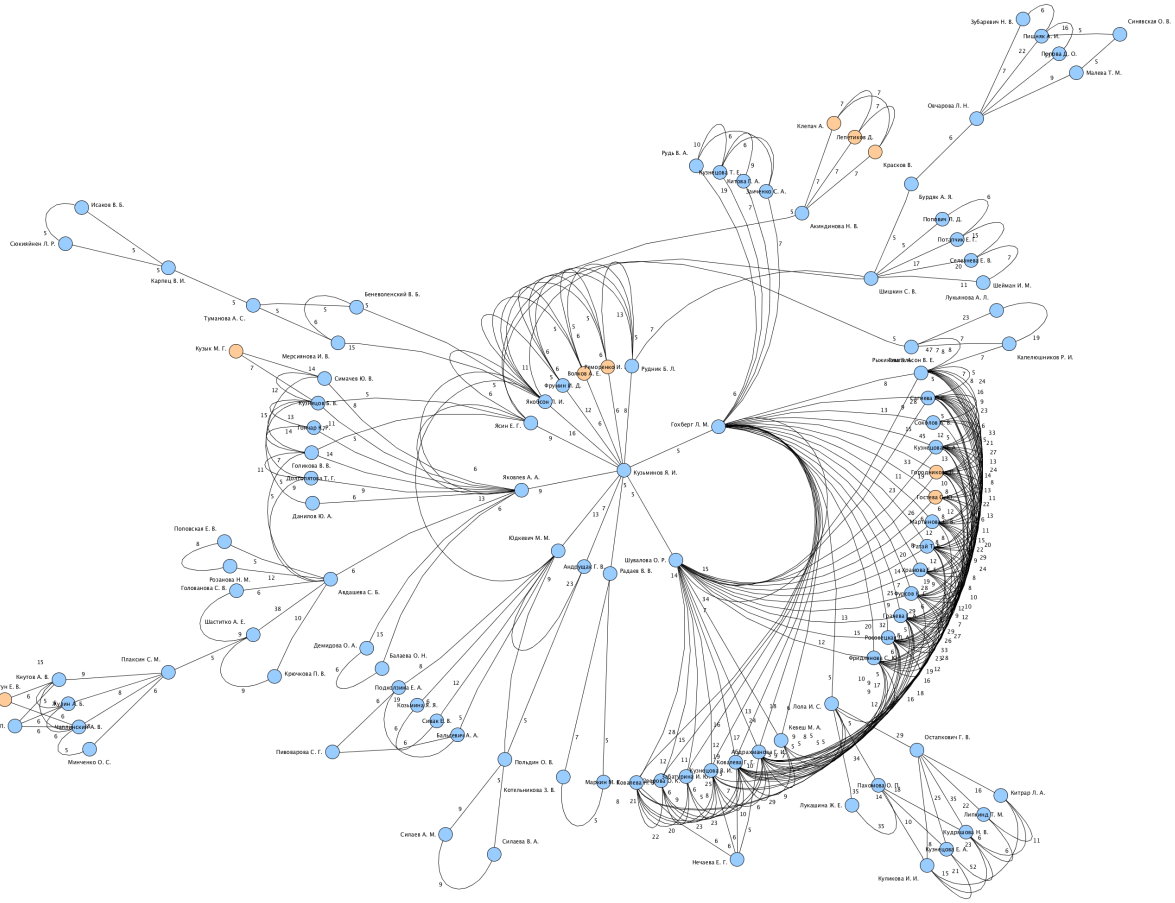

Figure 1: Visualizing co-authorship network for understanding administrative structure solely based on co-authorship network topology (the names are presented in original language).

many non-technician departments is quite low (see Table 1). That is why it is important to predict the core actors in research collaboration, which could improve the current situation and provide better research publications according to a certain measure. We choose quartiles in corresponding scientific indexing databases as such a metrics for the staff of the Higher School of Economics, we used the quality metric as a quartile (Q1, Q2, Q3, Q4, other), which is assigned to a journal or conference where the article is published and indexed in scientific libraries according to specific subject category (there may be several subject categories assigned to one journal). Quartile Q1 is assigned to the leading journals and conferences, then goes Q2, and after that point are usually journals with low-impact factor, which are not highly appreciated by research and scientommetrics communities. The information about journals and conferences has been taken from the SJR Scimago databases [23, 25] and InCites [51]. These databases contain information about publications that are indexed in Scopus and Web of Science Core Collection, respectively. Additionally, we disambiguate Russian journals, which are indexed in Scopus and contained English version with standalone translation [16]. We used specific Python libraries in order to imitate browser user activity in commercial web-services using Selenium, BeautifulSoup, urllib packages.

The research interests based on categories were added from [55]. The Web of Science journal research areas were then categorized as one or two of Scopus Subject Categories. In addition to quartiles, the average percent of journal self-citation for three years was collected. Such a metrics according to HSE rules make the level of journal lower due to its close community structure if self-citation becomes greater than 30%. For specific category “miscellaneous”, we studied the influence of this category on the change in the mean value of quartile among all categories for corresponding journal. We obtained that this category could be omitted as soon as it does not present the relevant research topics except the case in which the journal has only “miscellaneous” categories. For all the other cases, including the latter one, we compute the journal quartile as average of quartiles relevant to topic modelling of the paper, abstract and keywords. Information about network actors, such as administrative unit position, declared author research interests and additional interests derived from topic modelling BigARTM system [12]. However, compared to [40] we choose to use only research

interests presented as a binarized vector of Scopus Subject Categories, filled with missing values from topic modelling of author research papers. This approach helps us to preserve unified model of research interests synchronized with journal indexing.

3.4 Actor Attributes and Similarity Scores for Pairs of Vertices

We consider the problem of finding authors with similar interests to a selected one as a foundation for collaboration search. 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 [36]. We use *Common neighbors*, *Jaccard's coefficient*, *Adamic-Adar*, *Graph distance*, *Preferential Attachment* similarity scores as baseline for network descriptors for pairs of nodes (see Table 2).

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 and publication activity descriptive statistics, we use additional content-based and graph-based features and corresponding similarity score.

Nodes of the graph correspond to authors and, hence, have binarized and numeric attributes, such as

- Staff Attributes
  - Relation with NRU HSE;
  - Full-time/part-time status;
  - Department affiliation;
  - Administrative, research or lecturer position;
  - Type of position (lecturer, professor, senior researcher, etc.)
- Network Attributes
  - Degree centrality;
  - Betweenness centrality;
  - Closeness centrality;
  - Clustering;
- Publication Activity
  - Number of publications in total;
  - Number of publications in total in Scopus;
  - Weighted sum of quartiles of the current publications;
  - Number of publications in three years;
  - Number of publications for three years in Scopus;
  - Number of publications for three years in Scopus Q1, Q2;
  - Number of publications total in Scopus Q1, Q2;
- Unified vector of Subject Categories as a union of all the categories corresponding to the author papers;
- node2vec embedding with tuned p, q hyper-parameters during validation step.

As for edge representation, we use two-sided approach. At first, we compute several similarity scores based on separate node feature representations and obtain exact numerical features. Secondly, similar to graph embedding edge representation, we use the operations for representing elements of Cartesian product of feature spaces as some symmetric function applied by each component of independent vectorized feature space (see Table 3) with operators inspired by [24]).

We computed the number of publications between two authors and their weighted sum of quartiles for each respective edge (for negative sampling we make it equal zero) and combined several feature representations for edge embedding as a number of similarity scores (compact representation) and the nodes embedding pairwise representation via operators from Table 3

- Multi-Graph Edge Metric **(A)**
  - Number of multiple edges connecting two authors;
  - Weighted sum of quartiles of corresponding papers;
- Network Similarity Score **(B)**
  - Common Neighbors;
  - Jaccard's Coefficient;
  - Adamic-Adar Score;
  - Preferential Attachment;
  - Graph Distance;
- Subject Category Features **(C1)** (and their pairwise vectorized representations **(C2)**)
  - Hamming Distance;
  - Cosine Similarity;
  - Common Neighbors;
  - Jaccard's Coefficient;
- Author Attribute Features **(D1)** (and their pairwise vectorized representations **(D2)**)
  - Pearson Correlation Coefficient;
  - Cosine Similarity;
  - Jaccard's Generalized Coefficient;
- Unified vector of Subject Categories as a union of all the categories corresponding to the author papers **(E)**
- Edge Embedding based on node2vec actor embeddings **(F)**

4 TRAINING MODEL

In link prediction problem for co-authorship network, we have to separate problems

(1) for a given network predict missing links as a future collaborations that could lead to good research papers via binary classification methods;
(2) based on historical data predict the future links and their weights according to the suggested publication quality measurement via regression models and neural networks with MSE loss function;

For the first task, we filter our dataset removing 40% of existing edges preserving connectivity property of the network with edges left as positive examples, while proceeding with negative sampling of the same size as the set of edges in the network. For the second task, we use 80% for training set, 10% for validating hyper-parameters of node2vec graph embedding and choosing the best operator for edge representation, and 10% of remaining edges are used as test data.

We consider machine learning based approaches for evaluating regression/classification models based on regularized linear/logistic regression, SVR/SVM, XGBoost and Neural Networks with fully-connected layers and ELU activators. We have compared training models for the following combinations of edge training data attributes:

(1) Actor-based
   (a) **(B)**;

Table 1: HSE Research Activity on 2015 year

Features \Departments	Education	Computer Science and Math	Economics	Social	Management	Humanities	Law	Media & Design
Number of Authors	19,38	19,45	14,15	14,74	12	20,24	12,41	11,36
Number of Publications	442,15	434,52	231,63	402,01	187,33	256,88	264,71	146,86
Average Publications per Author from Department	27,66	26,55	21,3	30,51	15,44	13,67	22,3	10,94
Average Department Authors per Publication	0,09	0,06	0,07	0,05	0,09	0,12	0,07	0,1
Average Publications per HSE Authors	1,71	1,89	1,61	1,76	1,63	1,63	1,56	1,25
Average HSE Authors per Publications	0,66	0,54	0,65	0,62	0,65	0,71	0,67	0,82
Papers indexed by Scopus	5,08	47	4,5	4,39	2,13	2,79	0,24	1
Papers indexed by Scopus per Authpr	0,21	2,14	0,36	0,29	0,16	0,16	0,03	0,08
Papers indexed by Scopus for the latter 3 years	3,77	13	2,71	2,12	1,41	1,7	0,24	0,57
Papers indexed by Scopus for the latter 3 years per Author	0,16	0,59	0,21	0,15	0,12	0,09	0,03	0,07
Papers in Q1,Q2 quartiles in Scopus	2,31	36,07	2,88	2	1,63	2,48	0	1
Papers in Q1,Q2 quartiles in Scopus per Author	0,12	1,64	0,22	0,13	0,12	0,13	0	0,03
Papers in Q1,Q2 quartiles in Scopus for the latter 3 years	1,69	9,41	1,4	0,91	1,06	1,7	0	0,57
Papers in Q1,Q2 quartiles in Scopus for the latter 3 years per Author	0,1	0,4	0,1	0,06	0,1	0,09	0	0,07
Average Co-author Number	12,91	13,1	12,71	15,98	9,29	8,21	14	8,89
Number of Connected Components	9	10,14	7,33	6,7	6,31	9,73	5,29	7,57
Size of the Greatest Connected Component	9,62	7,43	6,65	8,28	5,97	9,91	7,47	4,43
Average Distance	1,51	1,43	1,43	1,55	1,35	1,52	1,42	1,19
Diameter	6,31	33,21	7,4	7,97	5,31	6,33	6,18	4,57
Clustering Coefficient	0,33	0,34	0,36	0,49	0,44	0,46	0,44	0,31
Density	0,2	0,12	0,15	0,19	0,16	0,14	0,29	0,09
Number of Lecturers	3,29	3,75	3,23	2,62	1,81	3,13	2,33	2,6
Number of Papers by Lecturers	3,3	4,37	2,41	3,7	3,43	4,73	6,17	2,03
Average Grade of Lecturers (0–30) as 15*(number of papers per 3 years)	25,71	19,99	16,09	17,91	15,69	18,85	19,06	18,25
Number of Senior Lecturers	3,3	4,76	2,98	4,71	5,43	5,18	7,05	2,03
Number of Papers by Senior Lecturers	25,71	21,81	39,93	22,8	25,1	20,64	21,79	18,25
Average Grade of Senior Lecturers (0–30) as 15*(number of papers per 3 years)	1,43	1,55	2,71	1,67	0,6	6,29	1,29	0,2
Number of Assistant Professors	25,71	21,81	19,93	22,8	25,1	20,64	21,79	13,25
Number of Papers by Assistant Professors	1,43	1,55	2,71	1,67	0,6	6,29	1,29	0,2
Average Grade of Assistant Professors (0–30) as 10*(number of papers per 3 years)	3,45	4,63	3,44	8	0,5	6,46	0,77	6
Number of Professors	1,86	2	2,41	2,19	1,89	6,36	2,8	1
Number of Papers by Professors	10,81	5,96	4,29	9,72	1,34	6,65	2,24	10
Average Grade of Professors (0–30) as 6*(number of papers per 3 years)	25,71	19,52	17,06	21,35	10,56	19,2	15,05	10

Table 2: Similarity score for a pair of nodes u and v with local neighborhoods N(u) and N(v) correspondingly, and for vectorized representations of two authors' attributes and research interests X and Y.

Similarity metric	Definition
Common Neighbors	$\lvert N(u) \cap N(v)\rvert$
Jaccard Coefficient	$\frac{\lvert N(u)\cap .N(v)\rvert}{\lvert N(u)\cup N(v)\rvert}$
Adamic-Adar Score	$\sum_{w\in N(u)\cap N(v)} \frac{1}{\ln \lvert N(w)\rvert}$
Preferential Attachment	$\lvert N(u)\rvert \cdot \lvert N(v)\rvert$
Graph Distance	length of shortest path between u and v
Metric score	$\frac{1}{1+\lvert\lvert x-y\rvert\rvert}$
Cosine score	$\frac{(x,y)}{\lvert\lvert x\rvert\rvert\lvert\lvert y\rvert\rvert}$
Pearson Coefficient	$\frac{cov(x,y)}{\sqrt{cov(x,x)}\cdot\sqrt{cov(y,y)}}$
Generalized Jaccard	$\frac{\sum min(x_i,y_i)}{\sum max(x_i,y_i)}$

Table 3: symmetric binary functions for computing vectorized (u, v)-edge representation based on node attribute embeddings $f(x)$ for ith component for $f(u, v)$

Symmetry Operator	Definition
Average	$\frac{f_i(u)+f_i(v)}{2}$
Hadamard	$f_i(u)\cdot f_i(v)$
Weighted-L_1	$\lvert f_i(u)-f_i(v)\rvert$
Weighted-L_2	$(f_i(u)-f_i(v))^2$
Neighbor Weighted-L_1	$\left\lvert \frac{\sum_{w\in N(u)\cup\{u\}} f_i(w)}{\lvert N(u)\rvert+1} - \frac{\sum_{t\in N(v)\cup\{v\}} f_i(t)}{\lvert N(v)\rvert+1}\right\rvert$
Neighbor Weighted-L_2	$\left(\frac{\sum_{w\in N(u)\cup\{u\}} f_i(w)}{\lvert N(u)\rvert+1} - \frac{\sum_{t\in N(v)\cup\{v\}} f_i(t)}{\lvert N(v)\rvert+1}\right)^2$

(b) **(B)+(C1)+(D1)+(E)**;

(2) Network-based

(a) **(F)**;

(b) **(F)+(C2)+(D2)**;

(3) Combined Model **(A)—(F)**

We measure AUC in order to compare the test results of trained models. The code for computing all the models with respect to classification and regression tasks, choosing proper edge embedding operator and tuning hyper-parameters of node embeddings will be uploaded on the paper Github http://github.com/makarovia/jcdl2018/.

We have obtained that combined approach of embedding network topology, author attributes and research subject categories leads to statistically significant improvement in the AUC for link prediction task for both, regression and classification tasks.

5 CONCLUSION

We have improved recommender systems [39, 40], based on various feature engineering techniques for co-authorship search, with a

new model involving graph embeddings. We have compared several machine learning solutions for link prediction task interpreted as binary classification problem and regression task for predicting the quality of possible joint publication. This system may be considered as a recommender system for searching 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 graph embedding component for extracting structural features from the network. 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 support novice researchers and increase their publishing activity or used to estimate the future collaboration of staff units. The combined deep learning approaches for integrating label, text, and network features for advanced graph embedding was left to the future work.

ACKNOWLEDGMENTS

This work has been funded by the Russian Academic Excellence Project '5-100'.

REFERENCES

[1] Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering Missing Links in Wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD '05)*. ACM, New York, NY, USA, 90–97. https://doi.org/10.1145/1134271.1134284

[2] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.

[3] Lars Backstrom and Jure Leskovec. 2011. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 635–644. https://doi.org/10.1145/1935826.1935914

[4] Mikhail Belkin and Partha Niyogi. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*. 585–591.

[5] Y. Bengio, A. Courville, and P. Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (Aug 2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[7] Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2015. GraRep: Learning Graph Representations with Global Structural Information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 891–900. https://doi.org/10.1145/2806416.2806512

[8] Bjarke Thorn Carstens, Mads Riis Jensen, Mathias Friis Spaniel, and Anders Hermansen. Vertex Similarity in Graphs using Feature Learning. (????).

[9] Nicola Cetorelli and Stavros Peristiani. 2013. Prestigious stock exchanges: A network analysis of international financial centers. *Journal of Banking & Finance* 37, 5 (2013), 1543–1551.

[10] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C. Aggarwal, and Thomas S. Huang. 2015. Heterogeneous Network Embedding via Deep Architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 119–128. https://doi.org/10.1145/2783258.2783296

[11] H. Chen, X. Li, and Z. Huang. 2005. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*. 141–142. https://doi.org/10.1145/1065385.1065415

[12] BigARTM contributors. 2016. BigARTM v0.8.2. https://doi.org/10.5281/zenodo.288960. (Dec. 2016). https://doi.org/10.5281/zenodo.288960

[13] NetworkX Developers. 2017. Link prediction algorithms. https://networkx.github.io/documentation/latest/reference/algorithms/link_prediction.html. (2017). [Online; accessed 17-January-2018].

[14] Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1, 3 (1936), 211–218.

[15] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 19, 1 (Jan 2007), 1–16. https://doi.org/10.1109/TKDE.2007.250581

[16] Elsevier. 2018. Russian WoS Journals. http://www.elsevierscience.ru/. (2018). [Online; accessed 9-January-2018].

[17] Andrew T. Fiore and Judith S. Donath. 2005. Homophily in Online Dating: When Do You Like Someone Like Yourself?. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1371–1374. https://doi.org/10.1145/1056808.1056919

[18] Michael Fire, Lena Tenenboim-Chekina, Rami Puzis, Ofrit Lesser, Lior Rokach, and Yuval Elovici. 2014. Computationally Efficient Link Prediction in a Variety of Social Networks. *ACM Trans. Intell. Syst. Technol.* 5, 1, Article 10 (Jan. 2014), 25 pages. https://doi.org/10.1145/2542182.2542192

[19] Valerio Freschi. 2009. A Graph-Based Semi-supervised Algorithm for Protein Function Prediction from Interaction Maps. In *Learning and Intelligent Optimization*, Thomas Stützle (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 249–258.

[20] Karl Pearson F.R.S. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572. https://doi.org/10.1080/14786440109462720 arXiv:https://doi.org/10.1080/14786440109462720

[21] Fei Gao, Katarzyna Musial, Colin Cooper, and Sophia Tsoka. 2015. Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming* 2015 (2015), 1.

[22] Sheng Gao, Ludovic Denoyer, and Patrick Gallinari. 2011. Temporal Link Prediction by Integrating Content and Structure Information. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 1169–1174. https://doi.org/10.1145/2063576.2063744

[23] Borja González-Pereira, Vicente P Guerrero-Bote, and Félix Moya-Anegón. 2010. A new approach to the metric of journalsâĂŹ scientific prestige: The SJR indicator. *Journal of informetrics* 4, 3 (2010), 379–391.

[24] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 855–864. https://doi.org/10.1145/2939672.2939754

[25] Vicente P Guerrero-Bote and Félix Moya-Anegón. 2012. A further step forward in measuring journalsâĂŹ scientific prestige: The SJR2 indicator. *Journal of Informetrics* 6, 4 (2012), 674–688.

[26] Mohammad Al Hasan and Mohammed J. Zaki. 2011. *A Survey of Link Prediction in Social Networks*. Springer US, Boston, MA, 243–275. https://doi.org/10.1007/978-1-4419-8462-3_9

[27] Xiao Huang, Jundong Li, and Xia Hu. 2017. Label Informed Attributed Network Embedding. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, New York, NY, USA, 731–739. https://doi.org/10.1145/3018661.3018667

[28] Gueorgi Kossinets and Duncan J Watts. 2009. Origins of homophily in an evolving social network. *American journal of sociology* 115, 2 (2009), 405–450.

[29] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[30] Eldon Y Li, Chien Hsiang Liao, and Hsiuju Rebecca Yen. 2013. Co-authorship networks and research impact: A social capital perspective. *Research Policy* 42, 9 (2013), 1515–1530.

[31] Xin Li and Hsinchun Chen. 2009. Recommendation As Link Prediction: A Graph Kernel-based Machine Learning Approach. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '09)*. ACM, New York, NY, USA, 213–216. https://doi.org/10.1145/1555400.1555433

[32] Xiaoyi Li, Nan Du, Hui Li, Kang Li, Jing Gao, and Aidong Zhang. *A Deep Learning Approach to Link Prediction in Dynamic Networks*. 289–297. https://doi.org/10.1137/1.9781611973440.33 arXiv:http://epubs.siam.org/doi/pdf/10.1137/1.9781611973440.33

[33] Yicong Liang, Qing Li, and Tieyun Qian. 2011. Finding relevant papers based on citation relations. In *IC on WAIM*. Springer, Springer, Berlin, 403–414.

[34] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2017. Attributed Social Network Embedding. *arXiv preprint arXiv:1705.04969* (2017).

[35] David Liben-Nowell and Jon Kleinberg. 2003. The Link Prediction Problem for Social Networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*. ACM, New York, NY, USA, 556–559. https://doi.org/10.1145/956863.956972

[36] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.

[37] Yan Liu and Zhenzhen Kou. 2007. Predicting Who Rated What in Large-scale Datasets. *SIGKDD Explor. Newsl.* 9, 2 (Dec. 2007), 62–65. https://doi.org/10.1145/1345448.1345462

[38] Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* 390, 6 (2011), 1150–1170.

[39] Bulanov O. Zhukov L.E. Makarov, I. 2017. Co-author Recommender System. In *Springer Proceedings in Mathematics and Statistic*. Springer, Berlin, 1–6. https://doi.org/10.1007/978-3-319-56829-4_18

[40] Ilya Makarov, Oleg Bulanov, Olga Gerasimova, Natalia Meshcheryakova, Ilia Karpov, and Leonid E. Zhukov. 2018. Scientific Matchmaker: Collaborator Recommender System. In *Analysis of Images, Social Networks and Texts*, Wil M.P. van der Aalst, Dmitry I. Ignatov, Michael Khachay, Sergei O. Kuznetsov, Victor Lempitsky, Irina A. Lomazova, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Andrey V. Savchenko, and Stanley Wasserman (Eds.). Springer International Publishing, Cham, 404–410.

[41] Bradley Malin, Edoardo Airoldi, and Kathleen M. Carley. 2005. A Network Analysis Model for Disambiguation of Names in Lists. *Comput. Math. Organ. Theory* 11, 2 (July 2005), 119–139. https://doi.org/10.1007/s10588-005-3940-3

[42] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[44] David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD IC*. ACM, NY, 500–509.

[45] Carlos Medicis Morel, Suzanne Jacob Serruya, Gerson Oliveira Penna, and Reinaldo Guimarães. 2009. 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 (2009), e501.

[46] Mark Newman. 2004. Who is the best connected scientist? A study of scientific coauthorship networks. *Complex networks* 1 (2004), 337–370.

[47] Mark EJ Newman. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of NAS* 101, suppl 1 (2004), 5200–5205.

[48] Shirui Pan, Jia Wu, Xingquan Zhu, Chengqi Zhang, and Yang Wang. 2016. Tri-party deep network representation. *Network* 11, 9 (2016), 12.

[49] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 701–710. https://doi.org/10.1145/2623330.2623732

[50] powered by HSE Portal. 2017. Publications of HSE. http://publications.hse.ru/en. (2017). [Online; accessed 9-May-2017].

[51] Thomson Reuters. 2018. InCites Journal Ranking. http://ipsciencehelp.thomsonreuters.com/inCites2Live/newIC2Group/newInCites.html. (2018). [Online; accessed 9-January-2018].

[52] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. 2007. Recent developments in exponential random graph (p*) models for social networks. *Social networks* 29, 2 (2007), 192–215.

[53] Sam T. Roweis and Lawrence K. Saul. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290, 5500 (2000), 2323–2326. https://doi.org/10.1126/science.290.5500.2323 arXiv:http://science.sciencemag.org/content/290/5500/2323.full.pdf

[54] Emre Sarigöl et al. 2014. Predicting scientific success based on coauthorship networks. *EPJ Data Science* 3, 1 (2014), 9.

[55] Scopus. 2016. Scopus subject Categories. http://www.scopind.com/2016/06/asjc-codelist.html. (2016). [Online; accessed 9-June-2016].

[56] Ajit P. Singh and Geoffrey J. Gordon. 2008. Relational Learning via Collective Matrix Factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 650–658. https://doi.org/10.1145/1401890.1401969

[57] Virinchi Srinivas and Pabitra Mitra. 2016. *Applications of Link Prediction*. Springer International Publishing, Cham, 57–61. https://doi.org/10.1007/978-3-319-28922-9_5

[58] Jiliang Tang and Huan Liu. 2012. Unsupervised Feature Selection for Linked Social Media Data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 904–912. https://doi.org/10.1145/2339530.2339673

[59] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1067–1077. https://doi.org/10.1145/2736277.2741093

[60] Lei Tang and Huan Liu. 2011. Leveraging social media networks for classification. *Data Mining and Knowledge Discovery* 23, 3 (01 Nov 2011), 447–478. https://doi.org/10.1007/s10618-010-0210-x

[61] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 5500 (2000), 2319–2323. https://doi.org/10.1126/science.290.5500.2319 arXiv:http://science.sciencemag.org/content/290/5500/2319.full.pdf

[62] Warren S Torgerson. 1958. Theory and methods of scaling. (1958).

[63] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. 2003. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology* 21, 6 (2003), 697–700.

[64] Theresa Velden and Carl Lagoze. 2009. Patterns of collaboration in co-authorship networks in chemistry-mesoscopic analysis and interpretation. In *12th International Conference on Scientometrics and Informetrics*. ISSI Society, Rio de Janeiro, 1–12.

[65] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, ACM, NY, 448–456.

[66] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural Deep Network Embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1225–1234. https://doi.org/10.1145/2939672.2939753

[67] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. 2015. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* 58, 1 (01 Jan 2015), 1–38. https://doi.org/10.1007/s11432-014-5237-y

[68] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, Cambridge.

[69] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. 2012. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*. Springer, 639–655.

[70] Hao Wu and Kristina Lerman. 2017. Network Vector: Distributed Representations of Networks with Global Context. *arXiv preprint arXiv:1709.02448* (2017).

[71] Erjia Yan and Ying Ding. 2009. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the IST Association* 60, 10 (2009), 2107–2118.

[72] S. Yan, D. Xu, B. Zhang, H. j. Zhang, Q. Yang, and S. Lin. 2007. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1 (Jan 2007), 40–51.

[73] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. 2015. Network Representation Learning with Rich Text Information.. In *IJCAI*. 2111–2117.

[74] Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Revisiting semi-supervised learning with graph embeddings. *arXiv preprint arXiv:1603.08861* (2016).

[75] Zhaoquan Yuan, Jitao Sang, Yan Liu, and Changsheng Xu. 2013. Latent Feature Learning in Social Media Network. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 253–262. https://doi.org/10.1145/2502081.2502284

[76] Jiawei Zhang, Philip S. Yu, and Zhi-Hua Zhou. 2014. Meta-path Based Multi-network Collective Link Prediction. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1286–1295. https://doi.org/10.1145/2623330.2623645

[77] Jianhan Zhu, Jun Hong, and John G. Hughes. 2002. Using Markov Models for Web Site Link Prediction. In *Proceedings of the Thirteenth ACM Conference on Hypertext and Hypermedia (HYPERTEXT '02)*. ACM, New York, NY, USA, 169–170. https://doi.org/10.1145/513338.513381