Dual network embedding for representing research interests in the link prediction problem on co-authorship networks

Ilya Makarov[1,2], **Olga Gerasimova**[1], **Pavel Sulimov**[1], **and Leonid E. Zhukov**[1]

[1]National Research University Higher School of Economics, School of Data Analysis 20 Myasnitskaya, Moscow, 101000 Russia

[2]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia

Corresponding author:

First Author[1]

Email address: iamakarov@hse.ru

ABSTRACT

We present a study on co-authorship network representation based on network embedding together with additional information on topic modeling of research papers and new edge embedding operator. We use link prediction model for constructing recommender system for searching collaborators with similar research interests. Extracting topics for each paper we construct keywords co-occurrence network and use its embedding for further generalizing author attributes. Standard graph feature engineering and network embedding methods were combined for constructing co-author recommender system formulated as link prediction problem and prediction of future graph structure. We evaluate our survey on the dataset containing temporal information on National Research University Higher School of Economics over twenty five years of research articles indexed in Russian Science Citation Index and Scopus. Our model of network representation shows better performance for stated binary classification tasks on several co-authorship networks.

Keywords: Co-authorship Networks, Co-occurrence Network, Recommender Systems, Network Embedding, Link Prediction, Machine Learning

1 INTRODUCTION

Nowadays, researchers struggle to find relevant scientific contributions among large variety of international conferences and journal articles. In order not to miss important improvements in various related fields of study, it is important to know the current state-of-art results while not reading all the papers tagged by research interests. One of the solutions is to search for the most "important" articles taking into account citation or centrality metrics of the paper and the authors with high influence on specific research field [29]. However, such method does not include collaborative patterns and previous history of research publications in co-authorship. It also does not measure the author professional skills and the ability to publish research results according to paper influence metrics, for example, journal impact factor.

We study the problem of finding collaborator depending on his/her research community, the quality of publications and structural patterns based on co-authorship network suggested by Newman in [42], [43]. Early unsupervised learning approaches for community detection in research networks were studied in [41, 9, 62, 56]. A review on social network analysis and network science can be found in [60, 4, 49].

We focus on the link prediction (LP) problem [31] in order to predict links in temporal networks and restore missing edges in complex networks constructed over noisy data. The LP algorithms can be used to extract the missing link or to detect abnormal interactions in a given graph, however, the most suitable case is to use LP for predicting the most probable persons for future collaboration, which we state as a problem of recommending a co-author using LP ranking [28]. Our model is designed to predict whether a

⁴⁵ pair of nodes in a network would have a connection. We can also predict the parameters of such an edge
⁴⁶ in terms of publication quality or number of collaborators corresponding to the predicted link [36, 37].

⁴⁷ In general, LP algorithms are widely used in several applications, such as web linking [2], search
⁴⁸ for real-world friends on social networks [3], citation recommender system for digital libraries [22]. A
⁴⁹ complete list of existing applied LP techniques can be found in [50].

⁵⁰ Recently, the improvement of machine learning techniques shifted the attention from manual feature
⁵¹ engineering to the vectorized information representation. Such methods have been successfully applied for
⁵² natural language processing and now are tested on network topology representation despite the fact that an
⁵³ arbitrary graph could not be described by its invariants. The approach of representing network vertices by
⁵⁴ a vector model depending on actor's neighborhood and similar actors is called graph (network) embedding
⁵⁵ [45]. The current progress of theoretical and practical results on network embeddings [45, 52, 10, 20]
⁵⁶ shows state-of-art performance on such problems as multi-class actor classification and link prediction.
⁵⁷ Although, the existing methods use not only structural equivalence and network homophily properties, but
⁵⁸ also the actor attributes, such as labels, texts, images, etc. A list of surveys on graph embedding models
⁵⁹ and applications can be found in [6, 14, 18, 12].

⁶⁰ In this paper, we study a co-authorship recommender system based on co-authorship network where
⁶¹ one or more of the coauthors belong to the National Research University Higher School of Economics
⁶² (NRU HSE) and the co-authored publications are only those indexed in Scopus. We use machine
⁶³ learning techniques to predict new edges based on network embeddings [20, 61] and edge characteristics
⁶⁴ obtained from author attributes. We compare our approach with state-of-the-art algorithms for link
⁶⁵ prediction problem using structural, attribute and combined feature space to evaluate the impact of the
⁶⁶ suggested approach on the binary classification task of predicting links in co-authorship network. Such
⁶⁷ obtained system could be applied for expert search, recommending collaborator or scientific adviser, and
⁶⁸ searching for relevant research publications similar to the work proposed in [35, 34]. In what follows, we
⁶⁹ describe solution to link prediction problem leading to evaluation of our recommender system based on
⁷⁰ co-authorship network embeddings and manually engineered features for HSE researchers.

2 RELATED WORK

2.1 Link Prediction

⁷³ The link prediction problem was stated in [31], in which Liben-Nowell and Kleinberg proposed using
⁷⁴ node proximity metrics. The evaluation of the proposed metrics for large co-authorship networks showed
⁷⁵ promising results for predicting future links based on network topology without any additional information
⁷⁶ on authors. Unsupervised structural learning was proposed in [51]. Gao et al. [17] presented temporal link
⁷⁷ prediction based on node proximity and its attributes determined by the content using matrix factorization.

⁷⁸ Two surveys on link prediction methods describe core approaches for feature engineering, Bayesian
⁷⁹ approach and dimensionality reduction were presented in [21, 33]. Survey on link prediction was published
⁸⁰ in [59].

⁸¹ The simplest baseline solution using network homophily is based on common neighbors or other
⁸² network similarity scores [31]. However, the Gao et al. showed in [16] that the similarity measures are
⁸³ not robust to the network global properties and, thus, could noise the prediction model with similarity
⁸⁴ scores only. The impact of the attribute-based formation in social networks was considered in [47, 39].
⁸⁵ All these observations require feature engineering depending on the domain.

⁸⁶ Graph-based recommender systems formulated via link prediction problem were suggested in
⁸⁷ [11, 32, 28]. In [26], Kossinets and Watts studied the effect of homophily in a university commu-
⁸⁸ nity. They considered temporal co-authorship network accompanied with author attributes and concluded
⁸⁹ the influence of not only structural proximity, but also author homophily for the social network structure.

⁹⁰ Another approach focusing on interdisciplinary collaboration inside the University was presented in
⁹¹ [13]. The authors used the existing co-authorship network and academic information for University of
⁹² Bristoll and proposed a new link prediction model for co-authorship prediction and recommendation.

⁹³ X. Kong et al. in [25] developed a scientific paper recommender system called VOPRec. However, in
⁹⁴ contrast to our work they constructed vector representation of research papers in citation networks. Their
⁹⁵ system uses both text information represented with word embedding to find papers of similar research
⁹⁶ interest and structural identity converted into vectors to find papers of similar network topology. To
⁹⁷ combine text and structural informations with the network, vector representation of article can be learned
⁹⁸ with network embedding.

2.2 Network Embedding

In general, knowledge retrieval and task-dependent feature extraction would require domain-specific expert to construct a real-value feature vector for nodes and edges representation. The quality of such an approach will be influenced by particular tasks and expert work, while not being scalable for large noisy networks. Recently, the theory of hidden representations has impacted on machine learning and artificial intelligence. It shifted the attention from manual feature engineering to defining loss function and then solving optimization task. The early works on network vectorized models were presented in Local Linear Embedding [48], IsoMAP [55], Laplacian Eigenmap [5], Spectral Clustering [54], MFA [63] and GraRep [7]. These works try to embed the networks into real-value vector space using several proximity metrics. However, development of representation learning for networks was in stagnation due to the non-robust and non-efficient machine learning methods of dimensionality reduction based on network matrix factorization or spectral decomposition. These methods were not applicable for large networks and noisy edge and attribute data providing low accuracy and having high time complexity of constructing embedding.

The modern methods of network embedding try to improve the performance on several typical machine learning tasks using conditional representation of a node based on its local and global neighborhood defined via random walking. The first-order and second-order nodes proximity were suggested in LINE [52] and SDNE [58] models. Generalizing this approach, DeepWalk [45] and node2vec [20] algorithms use Skip-gram model [40] based on simulation of breadth-first sampling and depth-first sampling. Although, in [8], Carstens et al. showed some drawbacks of node2vec [20] graph embedding, it still remained competitive structural-only embedding for representing both, homophily and structural equivalence in the network. Its generalization on global network representation learning from [61] shows comparable results with the original model.

Several works cover the node attributes, such as label and text content (see TADW [64], LANE [23]). In TriDNR paper [44], Pan et al. proposed to separately learn structural embedding from DeepWalk [45] and content embedding via Doc2Vec [27]. On the contrary, ASNE [30] learns combined representations for structural and node attribute representation using end-to-end neural network.

We focus on graph embedding and feature engineering methods applied to a link prediction task on a particular network, consisting of HSE researchers co-authored at least one paper with additional attributes representing authors. Using network features only fails to include the information about actors obtained from the other sources, thus decreasing efficiency of network embeddings. We aim to include information on feature space of author's research interests using data from the Scopus digital library containing manually input and automatically selected keywords for each research article. Based on this information we constructed keywords co-occurrence network and consider its embedding for further generalizing author attributes.

3 DATASET DESCRIPTION AND PREPROCESSING

We use NRU HSE portal [46] containing information on research papers co-authored by at least one HSE researcher, which were later uploaded to the portal by one of co-authors. The HSE database contains information on over 7000 HSE researchers published over 31000 research papers. The portal site contains web interface for the researchers to extract metadata of publications for a given time period and could be used by external researchers. The database records contain information on title, list of authors, keywords, abstract, year and place, journal/conference and publishing agency, and indexing flags for Scopus, Web of Science (WoS) Core Collection and Russian Science Citation Index (RSCI).

Unfortunately, the database has no interface for managing bibliography databases and has no integration with synchronizing of indexing digital libraries compared to Scholar Google or personal researcher profile management servises such as ResearcherID or Orcid. As a consequence, a large amount of noisy data occurs due to such problems as author name ambiguity or incorrect/incomplete information on the publications.

In order to resolve the ambiguity we considered standard disambiguation approaches for predicting necessity to merge authors. We used Levenshtein distance (useful for records with one-two error letters) for abbreviated Author Last and First names and then validated by two thresholds whether to merge two authors with the same abbreviation in the database based on cosine similarity and common neighbors metrics. The threshold values have been found manually via validation on small labeled set of ambiguous
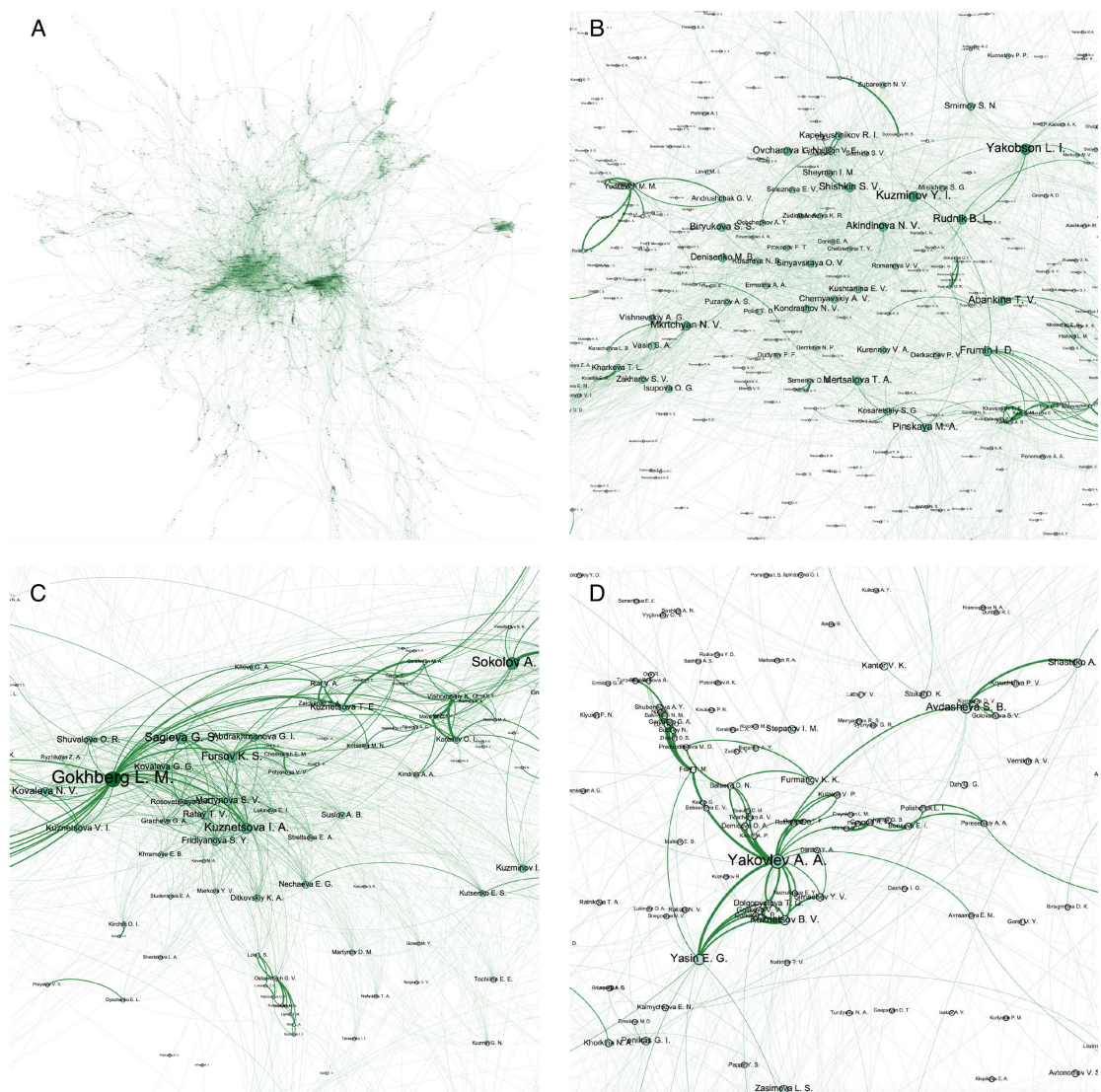
Figure 1. Visualization of HSE co-authorship network and its subgraphs

152 records. The number of authors with ambiguous writing does not exceed 2% of the whole database. We
153 have also removed all non-HSE authors due to lack of information on their publications in HSE dataset.

154   We also retrieved Scopus database of research papers co-authored by researchers from NRU HSE
155 and indexed by Scopus [15]. The database contains information on paper author list, document title, year,
156 source title, volume, issue, pages, source and document type, DOI, author keywords, index keywords. We
157 also added the information on research interests based on Scopus subject categories for the journals, in
158 which authors have published their articles. We manually inputed the research interest list according to
159 RSCI categorization in order to fill the lack of keywords and attributes for the papers.

160   We then stated the problem of indexing author research interests in terms of keywords attached to
161 paper description in both databases, and retrieved from HSE dataset using BigARTM [57] topic modeling
162 framework. For Scopus dataset we use already prepared by the service automatically chosen keywords
163 together with manually input by authors list of keywords. We also uses additional keywords written in
164 terms of subject categories of journals and proceedings according to the indexing in Scopus and Web of
165 Science research paper libraries.

166   These two datasets (HSE, Scopus) have common papers, however HSE dataset contains many noisy
167 data and, unfortunately, low-level publications not indexed outside RSCI, while Scopus contains precise

168 information on 25% number of papers and exact research interest representation while lacking weak
169 connections on authors written only poor-quality papers.

170 We visualized the HSE network with author names as node labels shown in the Figure 1 while
171 visualizing edge width as the cumulative quantity of joint publications based on papers' quartiles and their
172 number. In particular, we plotted the whole network structure (Fig. 1A) and zoomed parts corresponding
173 dense communities with the most influential persons from our university such as rector Kuzminov Y.I.
174 (Fig. 1B), first vice-rector responsible for science Gokhberg L.M. (Fig. 1C), and university research
175 supervisor Yasin E.G. (Fig. 1D). It is easy to see that rector co-authors are people responsible for core
176 direction of university development and vice rectors realizing university strategy in research, education,
177 government, expert and company areas of collaboration. On the other hand from dense subgraphs one
178 can find exact matches of university staff departments, such as research institutes, such as Institute for
179 Industrial and Market Studies headed by Yakovlev A.A. or Institute for Statistical Studies and Economics
180 of Knowledge headed by L.M. Gokhberg.

181 We show that network could visualize the most important administrative staff units and their heads
182 thus giving insight on the connection of publication activity and administrative structure of HSE university.

183 4 FEATURE ENGINEERING

184 Our new idea was to obtain additional edge attributes that were embed based on keywords network as a
185 part of model evaluation. We constructed the network of stemmed keywords co-occurrence. To construct
186 this network we used the principle that two nodes were connected if corresponding keywords occurred
187 in the same paper. For a given list of keywords we built standard node2vec embedding [20]. Next, for
188 each author the most frequent and relevant keyword was defined, and its embedding was used as node
189 additional feature vector for our LP tasks.

190 We considered the problem of finding authors with similar interests to a selected one as collaboration
191 search problem. In terms of social network analysis, we studied the problem of recommending similar
192 author as LP problem. We operate with authors similarity and use similarity scores described in [31] as
193 baseline for network descriptors for pairs of nodes presented in Table 1.

Table 1. Similarity score for a pair of nodes u and v with local neighborhoods N(u) and N(v) correspondingly, and for vectors corresponding to two authors research interests X and Y.

SIMILARITY METRIC	DEFINITION
COMMON NEIGHBORS	$|N(u) \cap N(v)|$
JACCARD COEFFICIENT	$\frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$
ADAMIC-ADAR SCORE	$\sum_{w \in N(u) \cap N(v)} \frac{1}{\ln|N(w)|}$
PREFERENTIAL ATTACHMENT	$|N(u)| \cdot |N(v)|$
GRAPH DISTANCE	LENGTH OF SHORTEST PATH BETWEEN u AND v
METRIC SCORE	$\frac{1}{1+||x-y||}$
COSINE SCORE	$\frac{(x,y)}{||x||||y||}$
PEARSON COEFFICIENT	$\frac{cov(x,y)}{\sqrt{cov(x,x)} \cdot \sqrt{cov(y,y)}}$
GENERALIZED JACCARD	$\frac{\sum min(x_i,y_i)}{\sum max(x_i,y_i)}$

194 So, we represented each node by vector model of author attributes using manually engineered features
195 such as HSE staff information and publication activity represented by centralities of co-authorship network

and descriptive statistics. We added graph embeddings for author research interests and node proximity and evaluated different combinations of models corresponding to node feature space representation.

5 LINK EMBEDDINGS

To use node2vec, we obtained the vector node representations. The node2vec embedding parameters were chosen via ROC-AUC optimization over embedding size with respect to different edge embedding operators. For edge embedding we applied specific component-wise functions representing edge to node embeddings for source and target nodes of a given edge. This model was suggested in [20], in which four functions for such edge embeddings were presented (see first four rows in Table 2). We leave an evaluation of the approaches from [1], which use bi-linear form learning from reduced by deep neural network node embedding, for the future work while also working on a new model of joint node-edge graph embedding, similar to [19]. Presented in the paper model suggests simple generalization of the idea of pooling first-order neighborhood of nodes while constructing edge embedding operator, which is much faster than dimensionality reduction approaches.

We evaluated our model on two additional functions involving not only edge source and target node representations, but also their neighborhood representations as average over all the nodes in first-order proximity. These measures were first presented in [38] but were not properly evaluated. The resulting list of link embeddings is presented in Table 2.

For each author, we also chose the most frequent keyword in the co-occurrence network and then constructed general embedding using node2vec with automatically chosen parameters.

Overall edge embedding contained several feature spaces described by the following List 1:

(1) Edge Embedding based on node2vec node embeddings

(2) Edge Embedding based on node2vec embedding for keywords co-occurrence network

(3) Network Similarity Scores (baselines in [31])

- Common Neighbors

- Jaccard's Coefficient

- Adamic-Adar Score

- Preferential Attachment

- Graph Distance

(4) Author Similarity Scores

- Cosine Similarity

- Common Neighbors

- Jaccard's Generalized Coefficient

- Pearson's Correlation Coefficient

- Metric Score

6 TRAINING MODEL

We consider machine learning models for binary classification task of whether for a given pair of nodes there will be a link connecting them based on previous or current co-authorship network. We compare Logistic Regression with Lasso regularization, Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine, in short SVM, models.

We use most common machine learning frameworks which we shortly describe below. *Logistic Lasso Regression* is a kind of linear model with a logit target function, in addition, Lasso regularization sets some

Table 2. Binary operators 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	$\dfrac{f_i(u) + f_i(v)}{2}$
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	$\left\lvert \dfrac{\sum_{w \in N(u) \cup \{u\}} f_i(w)}{\lvert N(u) \rvert + 1} - \dfrac{\sum_{t \in N(v) \cup \{v\}} f_i(t)}{\lvert N(v) \rvert + 1} \right\rvert$
NEIGHBOR WEIGHTED-L_2	$\left(\dfrac{\sum_{w \in N(u) \cup \{u\}} f_i(w)}{\lvert N(u) \rvert + 1} - \dfrac{\sum_{t \in N(v) \cup \{v\}} f_i(t)}{\lvert N(v) \rvert + 1} \right)^2$

coefficients to zero, effectively choosing a simpler model that reduces number of coefficients. *Random Forest* and *Gradient Boosting* are an ensemble learning methods. The former operates by constructing a multitude of decision trees, the latter combines several weak models building the model in a stage-wise fashion and generalizes them by allowing optimization of an arbitrary differentiable loss function. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. *SVM* constructs a hyperplane that has the largest distance to the nearest training-data point of any class when it achieves a good classification.

We use standard classification performance metrics for evaluating quality such as Precision, Accuracy, F1-score (micro, macro), Log-Loss, ROC-AUC. Next, we shortly define them.

Precision is a measure that tells us what proportion of publications that we diagnosed as existing, actually had existed. *Accuracy* in classification problems is the number of correct predictions made by the model over all kinds predictions made. *Recall* is a measure that tells us what proportion of publications that actually had existed was diagnosed by the algorithm as existing. To balance these metrics, *F1-score* is calculated as a weighted mean of the Precision and Recall. *Micro-averaged* metrics are usually more useful if the classes distribution is uneven by calculating metrics globally to count true and false predictions, *macro-averaged* are used when we want to evaluate systems performance across on different classes by calculating metrics for each label. *Log-loss*, or logarithmic loss, is a 'soft' measurement of Accuracy that integrates the idea of probabilistic confidence. In binary classification Log-loss can be calculated as $-\frac{1}{N} \cdot \sum_{i=1}^{N} (y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i))$, where y_i is a binary indicator (0 or 1) checking the correctness of classification for observation i, p_i is a predicted probability observation i for a given class. Log-Loss measures the unpredictability of the 'extra noise' that comes from using a predictor as opposed to the true labels. *ROC-AUC*, aka Area Under the Receiver Operating Characteristic Curve, is equal to the probability that a classifier will rank a randomly chosen existing edge higher than a randomly chosen non-existing one. ROC curves typically feature rate of true predicted existing publications on the Y axis, and rate of false predicted existing publications on the X axis. The larger area under the curve (AUC) is usually better. All metrics were averaged using 5-fold cross validation with 5 negative sampling trials for a fixed train set, which we describe below.

In our link prediction problem for co-authorship network, we have two possible formalizations of predicting links. We consider either temporal network structure using information from the previous years to predict links corresponding the current year or the whole network to predict missing links. For the first task, we use combined HSE+Scopus dataset of 2015 publications and learn to predict papers appearing at 2016 year. We test our model shifting years on one year ahead and evaluate our predictions for 2017 year based on publications until 2016 year. For the second task, we remove 50% of existing edges preserving connectivity property from our dataset and add negative sampling of the same size as a number of edges left in order to balance classes for classification problem.

In Table 3 for future links prediction task, we compare chosen predictive models fixing one Neighbor

Table 3. Comparing machine learning models based on Neighbor Weighted-L2 link embedding applied to future links prediction on Scopus dataset

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
LOGISTIC REGRESSION	0.382	0.372	0.483	0.471	2.089	0.534
RANDOM FOREST	0.622	0.671	0.652	0.641	0.893	0.725
GRADIENT BOOSTING	0.487	0.521	0.527	0.582	1.275	0.631
SVM	**0.712**	**0.784**	**0.761**	**0.754**	**0.634**	**0.816**

Table 4. Comparing link embeddings for future links prediction on Scopus dataset

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
AVERAGE	0.436	0.578	0.661	0.589	1.341	0.599
HADAMARD	0.613	0.654	0.657	0.628	1.125	0.634
WEIGHTED-L1	0.645	0.678	0.674	0.632	0.979	0.723
WEIGHTED-L2	0.672	0.682	0.688	0.637	0.915	0.742
NEIGHBOR WEIGHTED-L1	0.644	0.692	0.701	0.696	0.832	0.783
NEIGHBOR WEIGHTED-L2	**0.712**	**0.784**	**0.761**	**0.754**	**0.634**	**0.816**

Weighted-L2 link operator to construct edge embeddings considered as model features. It is interesting to see that XGBoost model get significantly overfitted while the best model appear to be Support Vector Machine.

In what follows, we aim to compare several link embedding metrics (see Table 2) for the best machine learning model. To evaluate our approach for the first task on Scopus dataset, we can see in Table 4 that suggested by authors new link embedding outperforms existing approaches by all the binary classification quality metrics.

As for the second task, we evaluate link prediction task over HSE and Scopus datasets in terms of predictive models and link embeddings. In Tables 5, 6 we could see that the SVM model outperforms the other model on HSE dataset but Random Forest get the best AUC for Scopus dataset being competitive with XGBoost and SVM models (SVM performs slightly better). This could happen due to sparse data in Scopus publications after removing 50% link information and lack of ensemble methods for choosing proper negative sampling for such a sparse dataset.

In Tables 7, 8 we could see that the suggested local proximity operator for link embedding that we call Neighbor Weighted-L2 link embedding outperforms all the other approaches for embedding edges based on node vector representations.

To choose the best predictive model and edge embedding operator, we consider several feature space combination based on List 1. The results of their comparison are shown in Table 9 for the first task and in Table 10 for the second task of LP (the original results appear in [38]).

We could see that adding embedding of author research interests as well as the author embedding itself play significant role in improving prediction quality for both tasks. When considering only structural embedding or node similarity features we obtained worse results in terms of all binary classification quality metrics. In both tasks, the combined approach with direct node similarity scores does not improve the quality of prediction overfitting the model on particular properties thus influencing the predictions for network with missing links. Makarov et al. [34] evaluate their recommender system on including research

Table 5. Comparing machine learning models based on Neighbor Weighted-L2 link embedding for link prediction problem on HSE dataset

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
LOGISTIC REGRESSION	0.421	0.462	0.502	0.472	1.873	0.583
RANDOM FOREST	0.836	0.871	0.774	0.831	0.193	0.888
GRADIENT BOOSTING	0.771	0.732	0.703	0.734	0.742	0.663
SVM	**0.823**	**0.845**	**0.782**	**0.812**	**0.273**	**0.828**

Table 6. Comparing machine learning models based on Neighbor Weighted-L2 link embedding for link prediction problem on Scopus dataset

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
LOGISTIC REGRESSION	0.482	0.491	0.522	0.563	1.452	0.613
RANDOM FOREST	0.812	**0.844**	**0.745**	**0.811**	**0.176**	**0.876**
GRADIENT BOOSTING	**0.852**	0.821	0.733	0.806	0.337	0.815
SVM	0.834	0.837	0.701	0.725	0.302	0.818

Table 7. Comparing link embeddings for link prediction problem on HSE dataset

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
AVERAGE	0.628	0.611	0.693	0.738	1.092	0.641
HADAMARD	0.721	0.728	0.673	0.687	0.703	0.771
WEIGHTED-L1	0.776	0.765	0.727	0.759	0.376	0.817
WEIGHTED-L2	0.786	0.782	0.751	0.782	0.355	0.827
NEIGHBOR WEIGHTED-L1	0.816	0.834	0.762	0.801	0.214	0.839
NEIGHBOR WEIGHTED-L2	**0.836**	**0.871**	**0.774**	**0.831**	**0.193**	**0.888**

Table 8. Comparing link embeddings for link prediction problem on Scopus dataset

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
AVERAGE	0.588	0.531	0.563	0.569	1.321	0.553
HADAMARD	0.672	0.637	0.611	0.632	0.784	0.626
WEIGHTED-L1	0.746	0.653	0.675	0.694	0.693	0.668
WEIGHTED-L2	0.786	0.771	0.705	0.707	0.597	0.774
NEIGHBOR WEIGHTED-L1	0.794	0.821	0.732	0.781	0.337	0.832
NEIGHBOR WEIGHTED-L2	**0.812**	**0.844**	**0.745**	**0.811**	**0.176**	**0.876**

Table 9. Prediction of publications for 2017 year based on $\leq$ 2016 years information

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
(1)	0,712	0,784	0,761	0,754	0,634	0,816
(3)	0,652	0,745	0,731	0,719	0,682	0,767
(1)+(2)	0,735	0,822	0,775	0,759	0,593	0,836
(1)+(3)	0,728	0,796	0,781	**0,789**	0,618	0,827
(1)+(4)	0,722	0,791	0,772	0,751	0,611	0,819
(3)+(4)	0,683	0,762	0,782	0,781	0,671	0,762
(1)+(2)+(3)	0,742	0,863	0,787	0,751	0,582	0,855
(1)+(2)+(4)	0,738	0,852	0,791	0,731	0,604	0,842
(1)+(2)+(3)+(4)	**0,798**	**0,866**	**0,793**	0,786	**0,573**	**0,878**

Table 10. Link prediction for 2017 year on Scopus dataset

	PRECISION	ACCURACY	F1-SCORE (MACRO)	F1-SCORE (MICRO)	LOGLOSS	ROC-AUC
(1)	0,767	0,831	0,703	0,783	0,236	0,859
(3)	0,731	0,822	0,698	0,778	0,244	0,813
(1)+(2)	0,811	0,843	0,726	0,805	0,216	0,864
(1)+(3)	0,763	0,821	0,703	0,775	0,231	0,839
(1)+(4)	0,772	0,833	0,715	0,794	0,223	0,861
(3)+(4)	0,747	0,824	0,719	0,791	0,241	0,825
§ (1)+(2)+(3)	0,808	0,835	0,724	0,807	0,193	0,866
(1)+(2)+(4)	**0,821**	0,842	0,733	**0,813**	0,203	0,872
(1)+(2)+(3)+(4)	0,812	**0,844**	**0,745**	0,811	**0,176**	**0,876**

299 interests based only on subject categories from the respective journals index in Scopus. It leads to worse
300 results for link prediction for the authors with small number of research publication, so they succeeded
301 only predicting so-called strong connections for authors writing at least three–five papers in co-authorship.
302 Our approach allows to work with arbitrary research interest representation thus making it possible to use
303 recommender system for novice researchers with small or zero connections in the network.

7 EXPERIMENTS FOR A LARGE NETWORK

305 To evaluate scalability of our results, we considered the link prediction task for the large network called
306 AMiner [53]. This network describes collaborations among the authors and contains 1,560,640 nodes and
307 4,258,615 edges.
308   For constructing node embeddings we used node2vec with model parameters $p, q = (1, 1)$, embedding
309 dimension $d = 128$, length of walk per node equaled $l = 60$, and number of walks per node equaled $n = 3$.
310 We decreased values of l and n in comparison with default values, because our computer terminated
311 process due to memory issues.
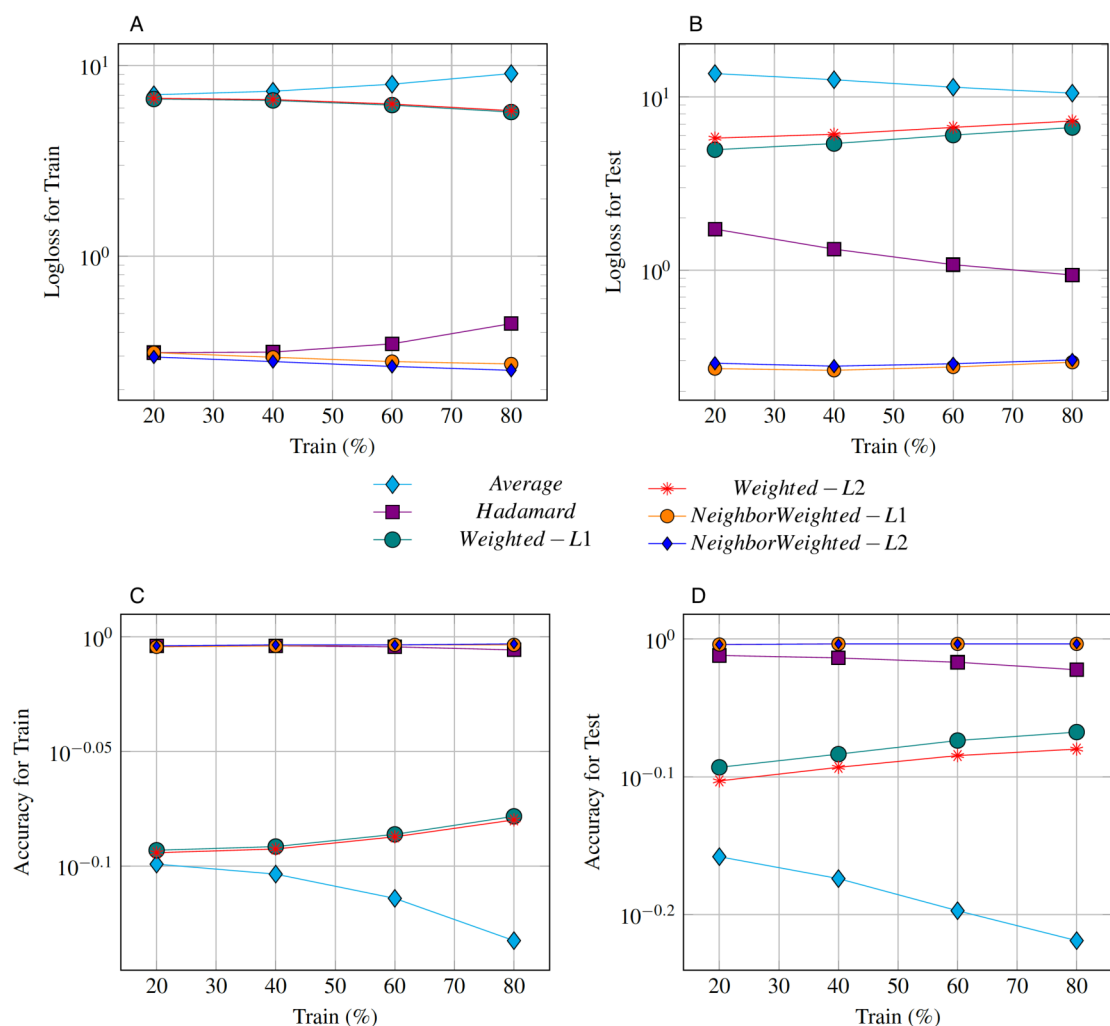
Figure 2. Low LogLoss and High Accuracy (plotted log-scale Y-axis) represent the best edge embedding. We show that NeighborWeighted-L_1 and NeighborWeighted-L_2 operators are on par with state-of-art Hadamard product presented in [20] and outperform it when train data size increases

312   We studied the impact of train/test split on different edge embeddings operators while fixing Logistic
313 Regression model for link prediction. We considered train set, consisting of 20%, 40% , 60% ,80% of

the graph edges while averaging binary classification quality metrics over 5 negative sampling providing negative examples for non-existent edges. We compared Logloss (Figs 2A and 2B) and Accuracy (Figs 2C and 2D) metrics computed for train and test sets using different edge embeddings. As a result, we obtained that Hadamard and NeighborWeighted-L_2 edge embedding operators represent highly accurate results while trained on sparse data from the original graph. Moreover, increasing the size of train data (which is the case for our temporal co-authorship network) Hadamard product becomes inferior to our NeighborWeighted-L_2 operator. In addition, NeighborWeighted-L_1 operator also showed greater performance in contrast to HSE dataset.

It was interesting to find out that overall performance of node2vec node embedding model produces very precise results for link prediction task, which may vary depending on graph as was shown in [20].

8 DISCUSSION

We have obtained that combined approach of embedding co-authorship and keywords co-occurrence networks, while preserving several author attributes leads to significant improvement in the classification quality metrics for predicting future links and link prediction tasks. However, we will continue to experiment with node2vec model. The (1) feature space from the List 1 is node2vec model with $d = 128$ (compared also on $d = 128$) and parameters $p, q = (1, 0.25)$ obtained by us via model fitting on a given d and logistic regression baseline model. The small q value shows that considering local proximity was more important than proximities of highest orders.

However we aim to further study this question taking into account modern methods of graph auto-encoders [24]. We are further working on a new embedding model called JONNE learning high quality node representations in tandem with edge representations. We proved the embeddings learned by JONNEE is almost always superior then of those learned by state-of-the-art models of all different types: matrix factorization based, sequence-based and deep learning based but the model has a certain drawback of longer training similar to matrix factorizations but less parallelizeable, thus giving us the dilemma to choose between quality and processing speed of suggested solution for edge embedding construction.

While link prediction task still remains hard problem for network analysis, its application to matching collaborators based on structural, attribute and content information shows promising results on applicability of graph-based recommender system predicting links in the co-authorship network and incorporating author research interests in collaborative patterns. We aim to generalize the model based on full-text extraction of research interests from collections of source documents and studying deep learning solutions for representing combined embedding of structural and content information for co-authorship networks.

The code for computing all the models with respect to classification evaluation, choosing proper edge embedding operator and tuning hyper-parameters of node embeddings will be uploaded on the paper Github `http://github.com/makarovia/jcdl2018/`, including HSE and Scopus datasets.

9 CONCLUSION

We have improved recommender systems [35, 34], based on choosing proper link embedding operator [38] and including research interest information presented as embedding of nodes in keywords co-occurrence network connecting keywords relating to a given research article. We have compared several machine learning models for future and missing link prediction problems interpreted as binary classification problem. The suggested by Makarov et al. in [38] edge embedding operator called Neighbor Weighted-L_2 (see Table 2) outperforms all the other edge embedding functions due to involving the neighborhood of the edge in the graph and was properly evaluated in this paper for both tasks. Among the machine learning models, SVM outperforms all the others except the LP problem on sparse Scopus dataset while XGBoost significantly get overfitted, however training SVM for large graphs is computationally hard.

Such constructed model may be considered as a recommender system for searching collaborators based on mutual research interests and publishing patterns. The recommender system demonstrates good results on predicting new collaborations between existing authors even if they have small number of data in co-authorship network due to availability of their research interests.

We are looking forward to the evaluation of our system for Universities, who have to deal with the problems of finding an expert based on text for evaluation, matchmaking for co-authored research papers with novice researchers, searching for collaborators on specific grant proposal or proper scientific advisers.

Focusing only on machine learning task is not suitable for real-world application involving social interactions, so we aim to implement framework with the possibility to manually add positive and negative preferences for collaboration recommendations thus providing useful service, which could be integrated in university business process of managing researchers' publication activity.

Our system may also be used for predicting the number of publications corresponding to a given administrative staff unit using network collaborative patterns and thus able to evaluate the efficiency of the authors or the whole staff. It also may be used for suggesting collaborations between separate staff units by considering combined network with staff units as vertices and weighted by the number of publications mutual connections between them. Evaluating such network for HSE university give us the picture that most popular faculties of Economics and Management have a lot of mutual connections due to many researchers working at these faculties, but limiting these connection only on Scopus publication leads to the influence of Computer Science and Engineering faculties which showed trend of computer science research in applied sciences. We leave for the future work consideration of the applicability of our system which suggests new University publication strategy based on collaboration patterns and invite the researchers to compare the existing solutions on HSE researchers dataset.

ACKNOWLEDGEMENTS

A part of the article is extended and revised version of presented oral talk at international conference AIST'18 and posters at ACM/IEEE JCDL'18, WebSci'18 and SunBelt'18. We thank all the colleagues from NRU HSE participated in discussion of this research.

REFERENCES

[1] Abu-El-Haija, S., Perozzi, B., and Al-Rfou, R. (2017). Learning edge representations via low-rank asymmetric projections. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1787–1796. ACM.

[2] Adafre, S. F. and de Rijke, M. (2005). Discovering missing links in wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 90–97, New York, NY, USA. ACM.

[3] Backstrom, L. and Leskovec, J. (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, pages 635–644, New York, NY, USA. ACM.

[4] Barabási, A.-L. and Pósfai, M. (2016). *Network science*. Cambridge university press.

[5] Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591.

[6] Cai, H., Zheng, V. W., and Chang, K. (2018). A comprehensive survey of graph embedding: problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering*.

[7] Cao, S., Lu, W., and Xu, Q. (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, pages 891–900, New York, NY, USA. ACM.

[8] Carstens, B. T., Jensen, M. R., Spaniel, M. F., and Hermansen, A. (2017). Vertex similarity in graphs using feature learning. [Online; accessed 07-06-2017].

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

[10] Chang, S., Han, W., Tang, J., Qi, G.-J., Aggarwal, C. C., and Huang, T. S. (2015). Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 119–128, New York, NY, USA. ACM.

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

[12] Chen, H., Perozzi, B., Al-Rfou, R., and Skiena, S. (2018). A tutorial on network embeddings. *arXiv preprint arXiv:1808.02590*.

[13] Cho, H. and Yu, Y. (2018). Link prediction for interdisciplinary collaboration via co-authorship network. *Social Network Analysis and Mining*, 8(1):25.

[14] Cui, P., Wang, X., Pei, J., and Zhu, W. (2018). A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*.

[15] Elsevier (2018). Scopus. `http://www.scopus.com/`. [Online; accessed 9-January-2018].

[16] Gao, F., Musial, K., Cooper, C., and Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, 2015:1.

[17] Gao, S., Denoyer, L., and Gallinari, P. (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, pages 1169–1174, New York, NY, USA. ACM.

[18] Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

[19] Goyal, P., Hosseinmardi, H., Ferrara, E., and Galstyan, A. (2018). Capturing edge attributes via network embedding. *arXiv preprint arXiv:1805.03280*.

[20] Grover, A. and Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA. ACM.

[21] Hasan, M. A. and Zaki, M. J. (2011). *A Survey of Link Prediction in Social Networks*, pages 243–275. Springer US, Boston, MA.

[22] He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 421–430, New York, NY, USA. ACM.

[23] Huang, X., Li, J., and Hu, X. (2017). Label informed attributed network embedding. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 731–739, New York, NY, USA. ACM.

[24] Kipf, T. N. and Welling, M. (2016). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

[25] Kong, X., Mao, M., Wang, W., Liu, J., and Xu, B. (2018). Voprec: Vector representation learning of papers with text information and structural identity for recommendation. *IEEE Transactions on Emerging Topics in Computing*.

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

[27] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

[28] Li, X. and Chen, H. (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, pages 213–216, New York, NY, USA. ACM.

[29] Liang, Y., Li, Q., and Qian, T. (2011). Finding relevant papers based on citation relations. In *IC on WAIM*, pages 403–414, Berlin. Springer, Springer.

[30] Liao, L., He, X., Zhang, H., and Chua, T.-S. (2017). Attributed social network embedding. *arXiv preprint arXiv:1705.04969*.

[31] Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology*, 58(7):1019–1031.

[32] Liu, Y. and Kou, Z. (2007). Predicting who rated what in large-scale datasets. *SIGKDD Explor. Newsl.*, 9(2):62–65.

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

[34] Makarov, I., Bulanov, O., Gerasimova, O., Meshcheryakova, N., Karpov, I., and Zhukov, L. E. (2018a). Scientific matchmaker: Collaborator recommender system. In *Analysis of Images, Social Networks and Texts*, pages 404–410, Cham. Springer International Publishing.

[35] Makarov, I., Bulanov, O., and Zhukov, L. (2017). Co-author recommender system. In *Springer Proceedings in Mathematics and Statistic*, pages 1–6, Berlin. Springer.

[36] Makarov, I., Gerasimova, O., Sulimov, P., Korovina, K., and Zhukov, L. (2019a). Joint node-edge network embedding for link prediction. In *Springer Proceedings in Mathematics and Statistic (to appear)*, pages 1–12, Berlin. Springer.

[37] Makarov, I., Gerasimova, O., Sulimov, P., and Zhukov, L. (2019b). Co-authorship network embedding and recommending collaborators via network embedding. In *Springer Proceedings in Mathematics*

472   *and Statistic (to appear)*, pages 1–6, Berlin. Springer.

473   [38] Makarov, I., Gerasimova, O., Sulimov, P., and Zhukov, L. E. (2018b). Recommending co-authorship
474   via network embeddings and feature engineering: The case of national research university higher school
475   of economics. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages
476   365–366. ACM.

477   [39] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social
478   networks. *Annual review of sociology*, 27(1):415–444.

479   [40] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations
480   of words and phrases and their compositionality. In *Advances in neural information processing systems*,
481   pages 3111–3119.

482   [41] Morel, C. M., Serruya, S. J., Penna, G. O., and Guimarães, R. (2009). Co-authorship network analysis:
483   a powerful tool for strategic planning of research, development and capacity building programs on
484   neglected diseases. *PLoS Negl Trop Dis*, 3(8):e501.

485   [42] Newman, M. E. (2004a). Coauthorship networks and patterns of scientific collaboration. *Proceedings
486   of NAS*, 101(suppl 1):5200–5205.

487   [43] Newman, M. E. (2004b). Who is the best connected scientist? a study of scientific coauthorship
488   networks. *Complex networks*, 1:337–370.

489   [44] Pan, S., Wu, J., Zhu, X., Zhang, C., and Wang, Y. (2016). Tri-party deep network representation.
490   *Network*, 11(9):12.

491   [45] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations.
492   In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data
493   Mining*, KDD '14, pages 701–710, New York, NY, USA. ACM.

494   [46] powered by HSE Portal (2017). Publications of hse. `http://publications.hse.ru/en`.
495   [Online; accessed 9-May-2017].

496   [47] Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007). Recent developments in
497   exponential random graph (p*) models for social networks. *Social networks*, 29(2):192–215.

498   [48] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding.
499   *Science*, 290(5500):2323–2326.

500   [49] Scott, J. (2017). *Social network analysis*. Sage.

501   [50] Srinivas, V. and Mitra, P. (2016). *Applications of Link Prediction*, pages 57–61. Springer International
502   Publishing, Cham.

503   [51] Tang, J. and Liu, H. (2012). Unsupervised feature selection for linked social media data. In
504   *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data
505   Mining*, KDD '12, pages 904–912, New York, NY, USA. ACM.

506   [52] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). Line: Large-scale information
507   network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW
508   '15, pages 1067–1077, Republic and Canton of Geneva, Switzerland. International World Wide Web
509   Conferences Steering Committee.

510   [53] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: Extraction and mining
511   of academic social networks. In *KDD'08*, pages 990–998.

512   [54] Tang, L. and Liu, H. (2011). Leveraging social media networks for classification. *Data Mining and
513   Knowledge Discovery*, 23(3):447–478.

514   [55] Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for
515   nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

516   [56] Velden, T. and Lagoze, C. (2009). Patterns of collaboration in co-authorship networks in chemistry-
517   mesoscopic analysis and interpretation. In *12th International Conference on Scientometrics and
518   Informetrics*, pages 1–12, Rio de Janeiro. ISSI Society.

519   [57] Vorontsov, K., Frei, O., Apishev, M., Romov, P., and Dudarenko, M. (2016). Bigartm v0.8.2.
520   `https://doi.org/10.5281/zenodo.288960`.

521   [58] Wang, D., Cui, P., and Zhu, W. (2016). Structural deep network embedding. In *Proceedings of the
522   22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16,
523   pages 1225–1234, New York, NY, USA. ACM.

524   [59] Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015). Link prediction in social networks: the state-of-the-art.
525   *Science China Information Sciences*, 58(1):1–38.

526   [60] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8.

Cambridge university press, Cambridge.

[61] Wu, H. and Lerman, K. (2017). Network vector: Distributed representations of networks with global context. *arXiv preprint arXiv:1709.02448*.

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

[63] Yan, S., Xu, D., Zhang, B., j. Zhang, H., Yang, Q., and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51.

[64] Yang, C., Liu, Z., Zhao, D., Sun, M., and Chang, E. Y. (2015). Network representation learning with rich text information. In *IJCAI*, pages 2111–2117.