Link Prediction Regression for Weighted Co-authorship Networks

Ilya Makarov[1,2(✉)] and Olga Gerasimova[1]

[1] National Research University Higher School of Economics, 3 Kochnovskiy Proezd, 125319 Moscow, Russia
{iamakarov,ogerasimova}@hse.ru
[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

Abstract. In this paper, we study the problem of predicting quantity of collaborations in co-authorship network. We formulated our task in terms of link prediction problem on weighted co-authorship network, formed by authors writing papers in co-authorship represented by edges between authors in the network. Our task is formulated as regression for edge weights, for which we use node2vec network embedding and new family of edge embedding operators. We evaluate our model on AMiner co-authorship network and showed that our model of network edge representation has better performance for stated regression link prediction task.

Keywords: Co-authorship networks · Recommender systems · Network embedding · Link prediction · Machine learning

1 Introduction

Co-authorship networks are powerful instrument to measure research trends, publication activity and study collaboration patterns in research community. There are various applications of co-authorship and citation networks to recommender systems for searching collaborators on research projects, reading relevant research papers, finding experts based on text of competition application. In order to meet the necessity to bring close researchers inside universities in developing countries and facilitate interdisciplinary research projects the concept of "Scientific matchmaking" was introduced in several research papers on co-authorship networks mining [39,49]. Early unsupervised learning

Sections 1, 2 and 3 on "Knowledge representation, discovery, and processing: a logic-based approach" were supported by the Russian Science Foundation under grant 17-11-01294 and performed at National Research University Higher School of Economics, Russia. Sections 4 and 5 on "Knowledge acquisition and representation for recommender systems" were prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'

I. Rojas et al. (Eds.): IWANN 2019, LNCS 11507, pp. 667–677, 2019.
https://doi.org/10.1007/978-3-030-20518-8_55

approaches for community detection and mining research networks were studied in [5,27,45,52,54,55].

The study on predicting individual collaborations was made by several research groups in Russia [35–37], Brazil [33], UK [11] and many other studies based on open datasets from DBLP and Scholar Google. On the state level the problem of improving research communities was studied at China [25].

In this paper, we study a co-authorship recommender system based on *node2vec* network embeddings [19] and edge characteristics obtained from author node embeddings. We compare our approach with state-of-the-art algorithms for link prediction problem using several edge embedding operators suggested in [19,40] and newly defined for the weighted link prediction (edge weight regression) problem state in the paper. Such obtained system could be applied recommending collaborators and estimating the outcome of such collaborations in temporal timeline. In what follows, we describe solution to link prediction problem leading to evaluation of our recommender system based on co-authorship network embeddings and evaluate its quality in regression and classification tasks for weighted network link prediction.

2 Related Works

2.1 Link Prediction

Network science approach to the problem of predicting collaborations results in the link prediction (LP) problem [28] for temporal networks and missing edges reconstruction in noisy network data. Basically, it is a method to apply standard machine learning framework for graph data considering feature space consisting of pairs of nodes and their features.

Link prediction models are applied in web linking [2], social dating services [4], paper recommender system for digital libraries [21]. A reader can found an up-to-date survey in [47]. Survey on link prediction was published in [53]. LP problem was specifically formulated in [28] based on nodes pairwise similarity measures. Approaches for link prediction include similarity based methods [3], maximum likelihood models [12], and probabilistic models [16,22]. In [51], authors suggesting unsupervised approach for LP problem. Authors of [14,15] suggested temporal link prediction based on matrix factorization technique and noise reduction in large networks. Attribute-based link formation in social networks was studied in [41,44], while deep learning approaches were presented in [6,29,56]. Heterogeneous graph link prediction for predicting links of certain semantic type was suggested in [31,32].

Two surveys on link prediction methods describe core approaches for feature engineering, Bayesian approach and dimensionality reduction were presented in [20,34]. Survey on link prediction was published in [53].

2.2 Collaborator Recommender System as Link Prediction Problem

Although link prediction can help detecting missing links in the current networks, or, on the contrary, abnormal links in fraud detection tasks, the most

of LP applications for social networks deals with predicting the most probable persons for future collaboration, which we state as a problem of co-authorship recommender system based on link prediction problem [10,26,30]. A survey on co-authorship and citation recommender systems may be found in [42].

We study the problem of recommending collaborator depending on researcher's co-authorship relations, the quality and quantity of publications, and structural patterns based in co-authorship network. We aim to measure the strength of such collaborations in terms of quantity of co-authored papers on Microsoft Academic Graph [46]. We also study the problem of estimating quality of such collaborations based on HSE university co-authorship network [23] with link weights aggregating numeric metrics based on impact-factors and quartiles of the journals for published research papers.

Our model is designed to predict whether a pair of nodes in a network would have certain number of connections and whether we could improve such a prediction in two-step process of predicting the collaboration itself [36–38] and further estimating its quantity/quality.

2.3 Graph Embedding

Recently, new methods of automated graph feature engineering and their applications to machine learning problems on graphs get attention of researchers. The network representation by adjusting each node a vector based on its neighborhood proximity is called network embedding. In general, knowledge representation requires task-dependent feature engineering in order to construct a real-value feature vector for nodes and edges representation. The quality of such an approach will be influenced by domain expert quality, particular task and noise in the data, which is hard to measure in large scale networks. Recently, the theory of constructing numeric embeddings has impacted on machine learning and artificial intelligence leading to creating new task-independent optimization tasks and loss functions instead of manual feature engineering.

New family of random walk based graph embeddings was suggested in such articles as [8,19,43,48]. The results of structural (without node attributes) graph embeddings already showed state-of-art performance on such problems as multi-class node classification and link prediction. A list of surveys on graph embedding models and applications can be found in [7,9,13,17].

3 Edge Feature Generating Based on Node Network Embedding

As we previously mentioned, we use node2vec method to construct node embeddings. Node2vec provides a flexible strategy to make a neighborhood sampling based on biased random walk, which smoothly combines two algorithms such as breadth-first and depth-first samplings. The biased random walk takes into account the second-order and even higher-order proximities. Applying a natural language processing methods such as vectorized technique word2vec to given sequences of neighborhood vertices generated by random walks, node2vec

receives node representations to optimize the occurrence probability of neighbor nodes based on the representation of a node.

The property of node2vec is ability to construct close embeddings for nodes belonging to the same network community or that are structurally equivalent. Node2vec has two random walk hyperparameters, p and q, that tune the random walk. The hyperparameter p regulates the chance that the walk revisits a node, while q controls situations, when the walk revisits a node's one-hop neighborhood. Changing settings of the model allows to find a reasonable compromise between learning embeddings to focus on community structures or local structural roles.

However, we are interested in receiving edges vector representation to form a feature space for link prediction task. 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 node2vec [19], in which four functions for such edge embeddings were presented (see first three rows in Table 1 for $\alpha = 1, 2$). We consider the simplest and fastest model of edge embedding construction not taking into account papers on joint node-edge graph embedding, such as [1,18]. 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 or graph neural networks.

In our previous research [40], we suggested to use another type of operators involving not only edge source and target node representations, but also their neighborhood representations as average over all the nodes in first-order proximity called "neighbour-weighted" L_α operator verifying parameter $\alpha = 1, 2$. The evaluation of such "vertex pooling" techniques was made in [38] for binary classification in link prediction problem.

Now for weighted link prediction we introduce new link embedding operators. Firstly, we consider Neighbor Weighted-L^0_α operator that differs from Neighbor Weighted-L_α just by not taking the vertex into account when pooling its neighborhood. Such an operator represent purely context based link embedding. Secondly, for weighted link prediction we change average pooling for vertex to weighted average sum of neighbour nodes embeddings multiplied by edge weight and divided by the sum of weights corresponding to the node. We call this Neighbor Double Weighted-L_α operator. For this case we do not place the upper index due to rear case of loops in the presented graph representing the papers written without co-authors. So, following the pooling idea, it is important not only to aggregate information from vertices, but also from their neighbourhood, because an edge appearance can be defined by second-order proximity too. The resulting list of link embeddings is presented in Table 1.

The core idea of original [19] edge embedding operators was to aggregate information for edge from incident nodes (average sum, Hadamard product) or choose L_α, $\alpha = 1, 2$ metric for the difference of embedding vectors to include the idea that similar (by adjacency matrix similarity), i.e., connected nodes should have close embeddings. Our idea follows both of these approaches but adds constraints that edge embedding is defined by second order proximity, meaning that neighbor nodes embeddings may be efficiently included in edge representation

Table 1. Binary operators for computing vectorized (u, v)-edge representation based on node attribute embeddings $f(x)$ for ith component for $f(u, v)$. Parameter $\alpha \geqslant 1$

Symmetry operator	Definition
Average (AVG)	$\frac{f_i(u)+f_i(v)}{2}$
Hadamard (MULT)	$f_i(u) \cdot f_i(v)$
Weighted-L_α (WL_α)	$\lvert f_i(u) - f_i(v)\rvert^\alpha$
Neighbor Weighted-L_α (NWL_α)	$\left\lvert \frac{\sum\limits_{w\in N(u)\cup\{u\}} f_i(w)}{\lvert N(u)\rvert+1} - \frac{\sum\limits_{t\in N(v)\cup\{v\}} f_i(t)}{\lvert N(v)\rvert+1} \right\rvert^\alpha$
Neighbor Weighted-L^0_α (NWL^0_α)	$\left\lvert \frac{\sum\limits_{w\in N(u)} f_i(w)}{\lvert N(u)\rvert} - \frac{\sum\limits_{t\in N(v)} f_i(t)}{\lvert N(v)\rvert} \right\rvert^\alpha$
Neighbor Double Weighted-L_α ($NDWL_\alpha$)	$\left\lvert \frac{\sum\limits_{w\in N(u)\cup\{u\}} f_i(w)\cdot\omega(u,w)}{\sum\limits_{p\in N(u)\cup\{u\}} \omega(u,p)} - \frac{\sum\limits_{t\in N(v)\cup\{v\}} f_i(t)\cdot\omega(v,t)}{\sum\limits_{s\in N(v)\cup\{v\}} \omega(v,s)} \right\rvert^\alpha$

construction. In what follows, we evaluate our previous and new edge embedding operators on link prediction problem.

4 Experiments

In order to state weighted link prediction problem we considered the link prediction task for the large network called AMiner [50] containing 4,258,615 collaborations among the 1,560,640 authors. We did not visualize it due to density of the data not allowing to make representative graph drawing. We work with weighted variant of Aminer graph, where an edge weight means the number of publications writing in collaboration between two authors corresponding to incident vertices for the given edge. Another existing weights interpretation that we do not consider in this article is related to publication quality, which is based on publication belonging to quartiles in scientific indexing databases (Q1–Q4, other), which are assigned to a journal or conference.

We construct *Linear Regression* model for predicting number of papers in AMiner graph written in co-authorship. We chose only this machine learning model due to existing time/memory limitations for large networks processing, but also due to high efficiency for our setting based on quality metrics.

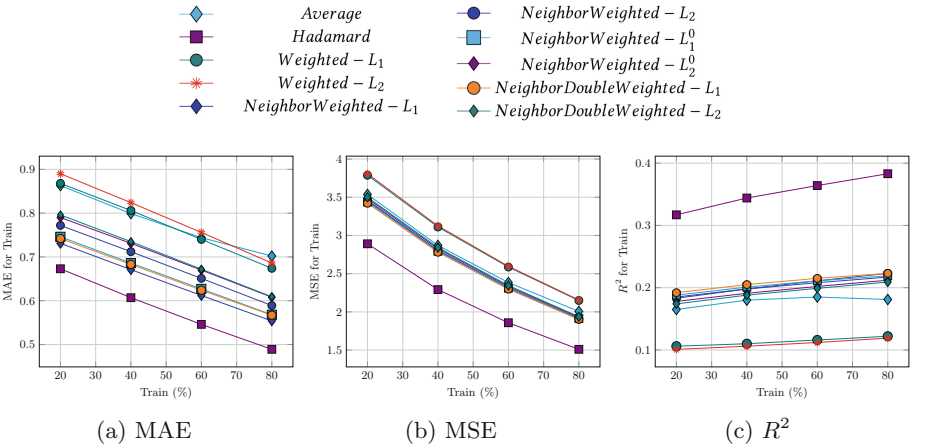
We studied the impact of train/test split on different edge embeddings operators while fixing the fastest Regression model for weighted link prediction. We considered train set, consisting of 20%, 40%, 60%, 80% of the graph edges while averaging binary classification quality metrics over 5 negative sampling providing negative examples for non-existent edges with zero weights.

We compared *Mean Absolute Error* (MAE) measuring the average of absolute values of the errors, *Mean Squared Error* (MSE) measuring the average of the squares of the errors, and *Coefficient of Determination* (R^2) measuring proportion of the variance in the predicted variable from the features computed for train and test sets using different edge embeddings. The results on train set

(Fig. 1) and test set (Fig. 2) are expressed in terms of MAE, MSE and R^2 metrics. The least values of MAE and MSE, and the highest values of R^2 represent the best models.

It turns out that Hadamard product outperforms all the others while the closest results was obtained by Neighbor Weighted-L_2 link embeddings. While train size increases all the Neighbor Weighted-L_α and Neighbor Double Weighted-L_α link embeddings achieve MAE less than 0.5, while simple Weighted-L_α and Average sum link embeddings showed the worst performance.

For our experiments We used node2vec network embedding [19] with random walks parameters $p, q = (1, 1)$, dimension of the embedding $d = 128$, length of walks $l = 60$, and number of walks per node equaled $n = 3$. The latter parameters are chosen quite small in order to fit to memory requirements. The node2vec embedding parameters were chosen via MSE optimization over embedding size with respect to Hadamard product edge embedding operators, which was stated to be the best for LP task in [19].

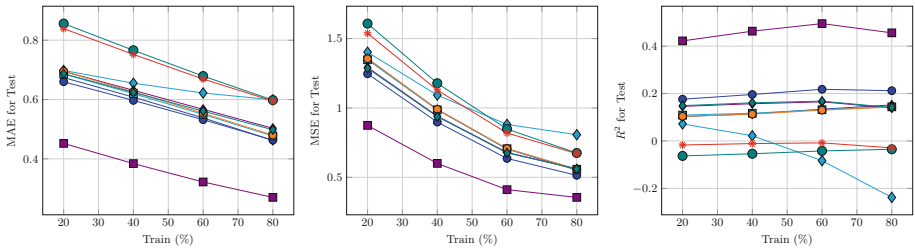


(a) MAE (b) MSE (c) R^2

Fig. 1. Linear regression results on train set depending on train size for different link embeddings

5 Discussion

As one could see, well performance of Hadamard product may be due the fact that such link embedding significantly reduce the dimension of link embedding feature space due to sparsity of embedding vector. However, its behavior on link prediction task formulated as binary classification showed bad learning ability of Hadamard product link embedding. In such a case pooling neighborhood must be done after encoding embedding to the lower space and then learning aggregated pooling function similar to [1]. From our experiments we could not deduce



(a) MAE (b) MSE (c) R^2

Fig. 2. Linear regression results on test set depending on train size for different link embeddings

whether average or weighted pooling significantly affect the performance to overcome Hadamard link embedding for regression task. On the opposite, classic link prediction task using binary classification based on edge embedding showed the best result on the Neighbor Weighted-L_α operators for HSE and AMiner graphs [38]. We aim to further study the problem of attributed network embeddings based on graph auto-encoders [24] and developed by authors embedding model called JONNEE learning joint node-edge representation of network.

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 project Github http://github.com/makarovia/jcdl2018/.

6 Conclusion

Scientific matchmaking is a new developing area in social network analysis. Its applications to analyzing researcher community and recommending new collaborations may be of interest to individuals, research groups, funding agencies and institutes. We previously showed that formulating the problem of collaborator search in terms of co-authorship network data showed high accuracy in predicting future collaborations for novice researchers [35,39] and analyzing research interaction inside university [36,37].

In this paper, we presented new link embedding operators similar to [38,40] including local proximity of source and target nodes for the edge. We verify the quality of such operators on weighted link prediction problem formulated as weight regression for given co-authorship networks. We considered AMiner graph with weights representing number of co-authored papers. Our approach of solving link prediction with weight regression presented MAE and MSE measures on test results less than 0.5.

Our experiments show that such constructed system may be considered as a recommender system for searching collaborators and simultaneously predictive model for estimating researcher's publishing and collaboration activity.

The recommender system demonstrates good results on predicting quantity and quality between existing and new collaborations.

References

1. Abu-El-Haija, S., Perozzi, B., Al-Rfou, R.: Learning edge representations via low-rank asymmetric projections. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1787–1796. ACM (2017)
2. Adafre, S.F., de Rijke, M.: Discovering missing links in wikipedia. In: Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD 2005, pp. 90–97. ACM, New York (2005). http://doi.acm.org/10.1145/1134271.1134284
3. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**(3), 211–230 (2003)
4. Backstrom, L., Leskovec, J.: 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 2011, pp. 635–644. ACM, New York (2011). http://doi.acm.org/10.1145/1935826.1935914
5. Barabási, A.L., Pósfai, M.: Network Science. Cambridge University Press, Cambridge (2016)
6. Berg, R.v.d., Kipf, T.N., Welling, M.: Graph convolutional matrix completion. arXiv preprint arXiv:1706.02263 (2017)
7. Cai, H., Zheng, V.W., Chang, K.: A comprehensive survey of graph embedding: problems, techniques and applications. IEEE Trans. Knowl. Data Eng. **30**, 1616–1637 (2018)
8. Chang, S., Han, W., Tang, J., Qi, G.J., Aggarwal, C.C., Huang, T.S.: Heterogeneous network embedding via deep architectures. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2015, pp. 119–128. ACM, New York (2015). http://doi.acm.org/10.1145/2783258.2783296
9. Chen, H., Perozzi, B., Al-Rfou, R., Skiena, S.: A tutorial on network embeddings. arXiv preprint arXiv:1808.02590 (2018)
10. Chen, H., Li, X., Huang, Z.: Link prediction approach to collaborative filtering. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005), pp. 141–142. IEEE (2005)
11. Cho, H., Yu, Y.: Link prediction for interdisciplinary collaboration via co-authorship network. Soc. Netw. Anal. Min. **8**(1), 25 (2018)
12. Clauset, A., Moore, C., Newman, M.E.: Hierarchical structure and the prediction of missing links in networks. Nature **453**(7191), 98 (2008)
13. Cui, P., Wang, X., Pei, J., Zhu, W.: A survey on network embedding. IEEE Trans. Knowl. Data Eng. **31**(5), 833–852 (2019)
14. Gao, F., Musial, K., Cooper, C., Tsoka, S.: Link prediction methods and their accuracy for different social networks and network metrics. Sci. Program. **2015**, 1 (2015)
15. Gao, S., Denoyer, L., Gallinari, P.: Temporal link prediction by integrating content and structure information. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 1169–1174. ACM, New York (2011). http://doi.acm.org/10.1145/2063576.2063744
16. Getoor, L., Taskar, B.: Statistical relational learning (2007)

17. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: a survey. Knowl.-Based Syst. **151**, 78–94 (2018)
18. Goyal, P., Hosseinmardi, H., Ferrara, E., Galstyan, A.: Capturing edge attributes via network embedding. arXiv preprint arXiv:1805.03280 (2018)
19. Grover, A., Leskovec, J.: Node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 855–864. ACM, New York (2016). http://doi.acm.org/10.1145/2939672.2939754
20. Hasan, M.A., Zaki, M.J.: A Survey of Link Prediction in Social Networks, pp. 243–275. Springer, Boston (2011). https://doi.org/10.1007/978-1-4419-8462-3_9
21. He, Q., Pei, J., Kifer, D., Mitra, P., Giles, L.: Context-aware citation recommendation. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 421–430. ACM, New York (2010). http://doi.acm.org/10.1145/1772690.1772734
22. Heckerman, D., Meek, C., Koller, D.: Probabilistic entity-relationship models, PRMS, and plate models. Introduction to statistical relational learning, pp. 201–238 (2007)
23. powered by HSE Portal: Publications of HSE (2017). http://publications.hse.ru/en. Accessed 9 May 2017
24. Kipf, T.N., Welling, M.: Variational graph auto-encoders. arXiv preprint arXiv:1611.07308 (2016)
25. Li, J., Xia, F., Wang, W., Chen, Z., Asabere, N.Y., Jiang, H.: ACREC: a co-authorship based random walk model for academic collaboration recommendation. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 1209–1214. ACM (2014)
26. Li, X., Chen, H.: 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 2009, pp. 213–216. ACM, New York (2009). http://doi.acm.org/10.1145/1555400.1555433
27. Liang, Y., Li, Q., Qian, T.: Finding relevant papers based on citation relations. In: Wang, H., Li, S., Oyama, S., Hu, X., Qian, T. (eds.) WAIM 2011. LNCS, vol. 6897, pp. 403–414. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23535-1_35
28. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Assoc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
29. Liu, F., Liu, B., Sun, C., Liu, M., Wang, X.: Deep learning approaches for link prediction in social network services. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013. LNCS, vol. 8227, pp. 425–432. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42042-9_53
30. Liu, Y., Kou, Z.: Predicting who rated what in large-scale datasets. SIGKDD Explor. Newsl. **9**(2), 62–65 (2007). https://doi.org/10.1145/1345448.1345462
31. Liu, Z., et al.: Semantic proximity search on heterogeneous graph by proximity embedding. In: AAAI, pp. 154–160 (2017)
32. Liu, Z., et al.: Distance-aware DAG embedding for proximity search on heterogeneous graphs. In: Thirty-Second AAAI Conference on Artificial Intelligence, pp. 2355–2362. AAAI (2018)
33. Lopes, G.R., Moro, M.M., Wives, L.K., de Oliveira, J.P.M.: Collaboration recommendation on academic social networks. In: Trujillo, J., et al. (eds.) ER 2010. LNCS, vol. 6413, pp. 190–199. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16385-2_24

34. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. Phys. A: Stat. Mech. Its Appl. **390**(6), 1150–1170 (2011)
35. Makarov, I., Bulanov, O., Zhukov, L.: Co-author recommender system. In: Kalyagin, V., Nikolaev, A., Pardalos, P., Prokopyev, O. (eds.) Models, Algorithms, and Technologies for Network Analysis. Springer Proceedings in Mathematics & Statistics, vol. 197, pp. 251–257. Springer, Berlin (2017). https://doi.org/10.1007/978-3-319-56829-4_18
36. Makarov, I., Gerasimova, O., Sulimov, P., Korovina, K., Zhukov, L.E.: Joint node-edge network embedding for link prediction. In: van der Aalst, W.M.P., et al. (eds.) AIST 2018. LNCS, vol. 11179, pp. 20–31. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-11027-7_3
37. Makarov, I., Gerasimova, O., Sulimov, P., Zhukov, L.E.: Co-authorship network embedding and recommending collaborators via network embedding. In: van der Aalst, W.M.P., et al. (eds.) AIST 2018. LNCS, vol. 11179, pp. 32–38. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-11027-7_4
38. Makarov, I., Gerasimova, O., Sulimov, P., Zhukov, L.: Dual network embedding for representing research interests in the link prediction problem on co-authorship networks. PeerJ Comput. Sci. **5**, e172 (2019)
39. Makarov, I., Bulanov, O., Gerasimova, O., Meshcheryakova, N., Karpov, I., Zhukov, L.E.: Scientific matchmaker: collaborator recommender system. In: van der Aalst, W.M.P., et al. (eds.) AIST 2017. LNCS, vol. 10716, pp. 404–410. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73013-4_37
40. Makarov, I., Gerasimova, O., Sulimov, P., Zhukov, L.E.: Recommending co-authorship via network embeddings and feature engineering: the case of national research university higher school of economics. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 365–366. ACM (2018)
41. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annu. Rev. Sociol. **27**(1), 415–444 (2001)
42. Ortega, F., Bobadilla, J., Gutiérrez, A., Hurtado, R., Li, X.: Artificial intelligence scientific documentation dataset for recommender systems. IEEE Access **6**, 48543–48555 (2018)
43. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2014, pp. 701–710. ACM, New York (2014). http://doi.acm.org/10.1145/2623330.2623732
44. Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P.: Recent developments in exponential random graph (p*) models for social networks. Soc. Netw. **29**(2), 192–215 (2007)
45. Scott, J.: Social Network Analysis. Sage, Thousand Oaks (2017)
46. Sinha, A., et al.: An overview of Microsoft Academic Service (MAS) and applications. In: Proceedings of the 24th international conference on world wide web, pp. 243–246. ACM (2015)
47. Srinivas, V., Mitra, P.: Applications of Link Prediction. In: Link Prediction in Social Networks. Springer International Publishing, Cham, pp. 57–61 (2016). https://doi.org/10.1007/978-3-319-28922-9_5
48. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015, pp. 1067–1077. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2015). https://doi.org/10.1145/2736277.2741093

49. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 990–998. ACM (2008)
50. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: KDD 2008, pp. 990–998 (2008)
51. Tang, J., Liu, H.: Unsupervised feature selection for linked social media data. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012, pp. 904–912. ACM, New York (2012). http://doi.acm.org/10.1145/2339530.2339673
52. Velden, T., Lagoze, C.: Patterns of collaboration in co-authorship networks in chemistry-mesoscopic analysis and interpretation. In: 12th International Conference on Scientometrics and Informetrics, pp. 1–12. ISSI Society, Rio de Janeiro (2009)
53. Wang, P., Xu, B., Wu, Y., Zhou, X.: Link prediction in social networks: the state-of-the-art. Sci. China Inf. Sci. **58**(1), 1–38 (2015). https://doi.org/10.1007/s11432-014-5237-y
54. Wasserman, S., Faust, K.: Social Network Analysis: Methods and applications, vol. 8. Cambridge University Press, Cambridge (1994)
55. Yan, E., Ding, Y.: Applying centrality measures to impact analysis: a coauthorship network analysis. J. IST Assoc. **60**(10), 2107–2118 (2009)
56. Zhai, S., Zhang, Z.: Dropout training of matrix factorization and autoencoder for link prediction in sparse graphs. In: Proceedings of the 2015 SIAM International Conference on Data Mining, pp. 451–459. SIAM (2015)