





On the Utilization of Structural and Textual Information of a Scientific Knowledge Graph to Discover Future Research Collaborations: A Link Prediction Perspective

Nikolaos Giarelis , Nikos Kanakaris , and Nikos Karacapilidis  

Industrial Management and Information Systems Lab, MEAD, University of Patras,
26504 Rio Patras, Greece

giarelis@ceid.upatras.gr, nkanakaris@upnet.gr, karacap@upatras.gr

Abstract. We consider the discovery of future research collaborations as a link prediction problem applied on scientific knowledge graphs. Our approach integrates into a single knowledge graph both structured and unstructured textual data through a novel representation of multiple scientific documents. The Neo4j graph database is used for the representation of the proposed scientific knowledge graph. For the implementation of our approach, we use the Python programming language and the scikit-learn ML library. We benchmark our approach against classical link prediction algorithms using accuracy, recall, and precision as our performance metrics. Our initial experimentations demonstrate a significant improvement of the accuracy of the future collaboration prediction task. The experimentations reported in this paper use the new COVID-19 Open Research Dataset.

Keywords: Link prediction · Research knowledge graphs · Natural language processing · Document representation · Future research collaborations

1 Introduction

In recent years, we have witnessed an increase in the adoption of graph-based approaches for predicting future research collaborations (Nathani et al. 2019; Vahdati et al. 2018). In these approaches, a collaboration between two researchers is generally denoted by a scientific article written by them (Ponomariov and Boardman 2016). Graph-based approaches (particularly those concerning knowledge graphs) build on concepts and methods from graph theory (e.g. node centrality, link prediction and node similarity measures) to discover hidden knowledge from the structural characteristics of the corresponding research graph (Wang et al. 2017). However, despite their broad adoption, existing graph-based approaches aiming to discover future research collaborations utilize only the structural characteristics of a research graph (Veira et al. 2019). In cases where unstructured textual data is available (e.g. graph nodes that correspond to scientific articles), existing approaches are incapable of simultaneously exploiting both the structural and the textual information of the graph.

To remedy the above weakness, this paper proposes the construction and utilization of a scientific knowledge graph where structured and unstructured data co-exist (e.g. document, author and word nodes). Building on our previous work, we represent the documents of a scientific graph as a *graph-of-docs* (Giarelis et al. 2020a; Giarelis et al. 2020b). This enables us to exploit both the structural and textual characteristics of a research graph towards building a novel link prediction algorithm for discovering future collaborations. The proposed approach uses the Neo4j graph database (<https://neo4j.com>) for the representation of the knowledge graph. For the implementation of our experiments, we use the Python programming language and the scikit-learn ML library (<https://scikit-learn.org>).

To evaluate the outcome of this paper, we benchmark our approach against different combinations of link prediction measures, which utilize only the structural information of a research graph. Our performance metrics include the accuracy, the precision, and the recall for each of the Machine Learning (ML) models considered. For our experiments, we use the COVID-19 Open Research Dataset (CORD-19). To examine whether our approach is affected by the size of the dataset (e.g. overfits or underfits), we extract and consider nine different well-balanced datasets. The experimental results show a significant improvement of the accuracy of the link prediction problem.

The remainder of the paper is organized as follows. Section 2 introduces background concepts and related work. Our approach is thoroughly presented in Sect. 3. Section 4 reports on the experiments carried out to evaluate the proposed approach. Limitations of our approach, future work directions and concluding remarks are outlined in Sect. 5.

2 Background Issues

For the discovery of future research collaborations, the proposed approach exploits a set of natural language processing (NLP), graph-based text representation, graph theory and knowledge graph techniques.

2.1 Graph Measures and Indices

Diverse graph measures and indices to capture knowledge related to the structural characteristics of a graph have been proposed in the literature (Vathy-Fogarassy and Abonyi 2013). Below, we mention a small subset of them, which is used in our approach.

The *Common Neighbors* measure, denoted by $CN(a, b)$, calculates the number of nodes that are common neighbors for a pair of nodes a and b (Li et al. 2018). It is defined as:

$$CN(a, b) = |\Gamma(a) \cap \Gamma(b)| \quad (1)$$

where $\Gamma(x)$ denotes the set of neighbors of a node x .

The *Total Neighbors* measure, denoted by $TN(a, b)$, takes into consideration all neighbors of a pair of nodes a and b (and not only the common ones as is the case in the previous measure). It is defined as:

$$TN(a, b) = |\Gamma(a) \cup \Gamma(b)| \quad (2)$$

The *Preferential Attachment* measure, denoted by $PA(a, b)$, calculates the product of the in-degree values of a pair of nodes a and b (Albert and Barabási 2001). This measure assumes that two highly connected nodes are far more likely to be connected in the future, in contrast to two loosely connected ones. This measure is defined as:

$$PA(a, b) = |\Gamma(a)| * |\Gamma(b)| \quad (3)$$

The *Adamic Adar* measure, denoted by $AA(a, b)$, calculates the sum of the inverse logarithm of the degree of the set of neighbors shared by a pair of nodes a and b (Adamic and Adar 2003). This measure assumes that nodes of a low degree are more likely to be influential in the future. It is defined as:

$$AA(a, b) = \sum_{c \in \Gamma(a) \cap \Gamma(b)} \left(\frac{1}{\log|c|} \right) \quad (4)$$

Finally, the *Jaccard Coefficient* index, denoted by $J(a, b)$, resembles the CN measure mentioned above; however, it differs slightly in that, for a pair of nodes a and b , it considers the amount of the intersection of their neighbor nodes over the union of them (Jaccard 1901). It is defined as:

$$J(a, b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{|\Gamma(a) \cup \Gamma(b)|} \quad (5)$$

2.2 Graph-Based Text Representations

The *graph-of-words* textual representation (Rousseau et al. 2013) represents each document of a corpus as a single graph. In particular, each graph node corresponds to a unique word of a document and each edge denotes the co-occurrence between two words within a sliding window of text. Rousseau et al. (2015) suggest that a window size of four seems to be the most appropriate value, in that it does not sacrifice either the performance or the accuracy of the ML models. Compared to the *bag-of-words* representation, it enables a more sophisticated feature engineering process due to the fact that it takes into consideration the co-occurrence between the terms. In any case, the limitations of the graph-of-words text representation are that: (i) it is unable to assess the importance of a word for a whole set of documents; (ii) it does not allow for representing multiple documents in a single graph, and (iii) it is not easily expandable to support more complicated data architectures.

To remedy the shortcomings of the graph-of-words representation, Giarelis et al. (2020b) have proposed the *graph-of-docs* representation, which depicts and elaborates multiple textual documents as a single graph. This last representation: (i) enables the investigation of the importance of a term into a whole corpus of documents, and (ii) it allows multiple node types to co-exist in the same graph, thus being easily expandable and adaptable to more complex data. In this paper, we utilize the graph-of-docs model to represent the textual data of a knowledge graph.

2.3 Related Work

As far as the discovery of future research collaborations using link prediction techniques is concerned, works closest to ours are those of Liben-Nowell and Kleinberg (2007), Sun et al. (2011), Guns and Rousseau (2014), Huang et al. (2008), and Yu et al. (2014). Specifically, Liben-Nowell and Kleinberg (2007) rely only on network topology aspects of a co-authors network, and the proximity of a pair of nodes to calculate the probability of future research collaborations between them. Sun et al. (2011) propose the use of structural properties to predict future research collaborations in heterogeneous bibliographic networks, where multiple types of nodes (e.g. venues, topics, papers, authors) and edges (e.g. publish, mention, write, cite, contain) co-exist. They exploit the relationships between the papers to improve the accuracy of their link prediction algorithm.

Guns and Rousseau (2014) recommend potential research collaborations using link prediction techniques and a random forest classifier. For each pair of nodes of a co-authorship network, they calculate a variety of topology-based measures such as Adamic Adar and Common Neighbors, and they combine them with location-based characteristics related to the authors. Hence, they propose future collaborations based on the location of the authors and their position on the co-authorship network. Huang et al. (2008) construct a co-authorship network for the Computer Science field that represents research collaborations from 1980 to 2005. They rely on classical statistical techniques and graph theory algorithms to describe the properties of the constructed co-authorship network. The dataset used contains 451,305 papers from 283,174 authors.

Yu et al. (2014) utilize link prediction algorithms to discover future research collaborations in medical co-authorship networks. For a given author, they attempt to identify potential collaborators that complement her as far as her skillset is concerned. They calculate common topological and structural measures for each pair of author nodes, including Adamic Adar, Common Neighbors and Preferential Attachment. ML models are used for the identification of possible future collaborations.

For a broader link prediction perspective, we refer to (Fire et al. 2011), (Julian and Lu 2016) and (Panagopoulos et al. 2017); these works describe approaches concerning the task of predicting possible relationship types between nodes (e.g. friendships in social networks).

3 Our Approach

Our approach first constructs a scientific knowledge graph that contains both structured and unstructured textual data. The integration of the unstructured textual data into the knowledge graph is accomplished through a graph-based text representation, namely *graph-of-docs* (see Sect. 2.2). Then, it employs graph measures and graph similarity techniques to extract features associated to both structural and textual information concerning the entities of a knowledge graph. Finally, it utilizes the produced features to build an ML model, which discovers future research collaborations by mapping the whole problem to a link prediction task. A detailed description of the abovementioned steps appears in (Giarelis et al. 2020a; Giarelis et al. 2020b).

3.1 The Scientific Knowledge Graph

Our knowledge graph allows diverse types of entities and relationships to co-exist in a the same graph data schema, including entity nodes with types such as 'Paper', 'Author', 'Laboratory', 'Location', 'Institution' and 'Word', and relationship edges with types such as 'is_similar', 'cites', 'writes', 'includes', 'connects', 'co_authors' and 'affiliates_with' (see Fig. 1).

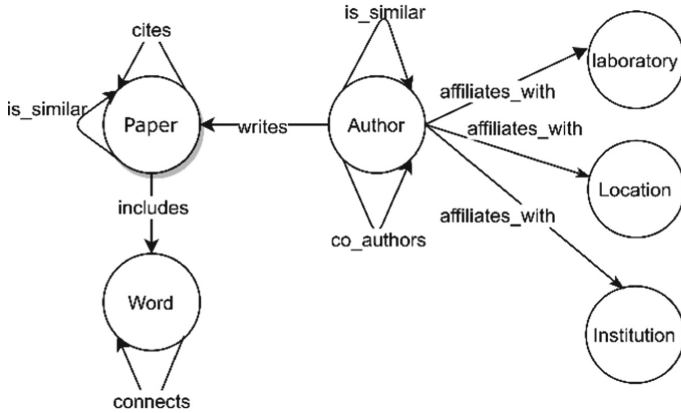


Fig. 1. The data schema of the scientific knowledge graph.

A 'Paper' entity represents a scientific paper or document. An 'Author' entity represents an author of a scientific paper or document. The 'Laboratory' entity represents the laboratory of an author. The 'Location' entity represents the location of a laboratory. The 'Institution' entity represents the institution of an author. Each 'Word' entity corresponds to a unique word of a scientific paper or document.

An 'includes' relationship connects a 'Word' with a 'Paper' entity. It marks the presence of a specific word to a certain paper. A 'connects' relationship is only applicable between two 'Word' entities and denotes their co-occurrence within a predefined sliding window of text. The subgraph constructed by the 'Word' and 'Paper' entities, as well as the 'includes', 'connects' and 'is_similar' relationships, corresponds to the graph-of-docs representation of the textual data of the available papers (see Fig. 2).

An 'is_similar' relationship links either a pair of 'Paper' or 'Author' nodes. In the former case, it denotes the graph similarity of the graph-of-docs representation of each paper. In the latter, it denotes the graph similarity between the graph-of-docs representations associated to the two authors. The subgraph that consists of the 'Author' entities and the 'is_similar' relationships corresponds to the authors similarity subgraph.

A 'cites' relationship links two 'Paper' nodes. A 'writes' relationship links an 'Author' with a 'Paper' entity. An 'affiliates_with' relationship

collaborations to the common binary classification problem. By using a binary classifier, we are able to predict the presence or the absence of a 'co_authors' relationship between two 'Author' entities, and thus build a link prediction algorithm for the discovery of future research collaborations. Available binary classifiers include logistic regression, k-nearest neighbors, linear support vector machines, decision tree, and neural networks (Aggarwal 2018).

4 Experiments

For the implementation and evaluation of our approach, we used the Python programming language and the scikit-learn ML library (<https://scikit-learn.org>). The Neo4j graph database (<https://neo4j.com>) has been utilized for the representation of the graph-of-docs and the corresponding knowledge graph. The full code, datasets, and evaluation results of our experiments are freely available at https://github.com/NC0DER/CORD19_GraphOfDocs.

4.1 CORD-19

The COVID-19 Open Research Dataset (CORD-19) (Wang et al. 2020) contains information about 63,000 research articles, related to COVID-19, SARS-CoV-2 and other similar coronaviruses. It is freely distributed from the Allen Institute for AI and Semantic Scholar (<https://www.semanticscholar.org/cord19>). The articles in CORD-19 have been collected from popular scientific repositories and publishing houses, including Elsevier, bioRxiv, medRxiv, World Health Organization (WHO) and PubMed Central (PMC). Each scientific article in CORD-19 has a list of specific attributes, namely 'citations', 'publish time', 'title', 'abstract' and 'authors', while the majority of the articles (51,000) also includes a 'full text' attribute. Undoubtedly, the CORD-19 dataset is a valuable source of knowledge as far as the COVID-19-related research is concerned; however, the fact that the majority of the data included is unstructured text renders a set of limitations in its processing. As advocated in the literature, the exploitation of a graph-based text representation in combination with a knowledge graph seems to be a promising step towards structuring this data (Veira et al. 2019; Wang et al. 2017; Wang et al. 2016). For the construction of our scientific knowledge graph, we utilize the 'abstract', 'authors' and 'publish time' attributes of each scientific article. We do not exploit the 'full text' attribute due to hardware limitations; however, we assume that the abstract of a paper consists a representative piece of its full text.

4.2 Experimental Setup

Selection of measures and metrics. To construct the authors similarity subgraph and to populate the edges of the 'Author'. 'is_similar' type, we use the Jaccard similarity index, since it deals only with the percentage of common set of words versus all words, ignoring their document frequency.

Construction of datasets for the link prediction problem. To test whether our approach performs well and does not overfit, regardless of the sample size of the dataset, we extract nine different datasets from the original one, corresponding to different volumes of papers (ranging from 1,536 to 63,023). For the creation of a sample creation, we utilize (i) the authors similarity subgraph, and (ii) the co-authors subgraph (i.e. the subgraph generated from the 'co_authors' edges; it is noted that edges also store the year of the first collaboration between authors, as a property). The features of a sample encapsulate either structural or textual characteristics of the whole knowledge graph (e.g. the similarity between the papers of two authors). Furthermore, each sample describes the relationship between two 'Author' nodes of the knowledge graph.

The features of a sample are analytically described in Table 1. Each of the nine datasets consists of a different number of randomly chosen samples. All datasets are balanced, in that the number of positive and negative samples are equal (see Table 2). To examine whether the features taken into account each time affect the efficiency of the ML models, we execute a set of experiments with different combinations of selected features (see Table 3). Finally, it is noted that the samples for the training subset are selected from an earlier instance in time of the co-authors subgraph, which is created from 'co_authors' edges first appeared within or before the year of 2013; respectively, the samples of the testing subset include 'co_authors' edges created after 2013. This separation in time ensures that we avoid any data leakage between the training and testing subsets (Liben-Nowell and Kleinberg 2007).

Table 1. A detailed explanation of the features of a sample. Each feature is associated to either a structural or a textual relationship between two given 'Author' nodes.

Feature	Description	Type
ademic_adar	The sum of the inverse logarithm of the degree of the set of common neighbor 'Author' nodes shared by a pair of nodes	Structural
common_neighbors	The number of neighbor 'Author' nodes that are common for a pair of 'Author' nodes	Structural
preferential_attachment	The product of the in-degree values of a pair of 'Author' nodes	Structural
total_neighbors	The total number of neighbor 'Author' nodes of a pair of 'Author' nodes	Structural
similarity	The textual similarity of the graph-of-docs graphs of two 'Author' nodes. The Jaccard index is used to calculate the similarity	Textual
label	The existence or absence of a 'co_authors' edge between two 'Author' nodes. A positive label (1) denotes the existence, whereas the absence is denoted by a negative label (0)	Class

Table 2. Number of samples (*lsamplesl*), number of positive (*lpositivel*) and negative (*lnegativel*) samples of the training and testing subsets of each dataset. A positive sample denotes the existence of a ‘*co_authors*’ edge between two ‘*Author*’ nodes, while a negative sample denotes the absence of such an edge.

	Training subset			Testing subset		
	<i>lsamplesl</i>	<i>lpositivel</i>	<i>lnegativel</i>	<i>lsamplesl</i>	<i>lpositivel</i>	<i>lnegativel</i>
Dataset 1	668	334	334	840	420	420
Dataset 2	858	429	429	1566	783	783
Dataset 3	1726	863	863	2636	1318	1318
Dataset 4	3346	1673	1673	7798	3899	3899
Dataset 5	5042	2521	2521	12976	6488	6488
Dataset 6	5296	2648	2648	16276	8138	8138
Dataset 7	6210	3105	3105	25900	12950	12950
Dataset 8	8578	4289	4289	34586	17293	17293
Dataset 9	13034	6517	6517	49236	24618	24618

Table 3. Combinations of features aiming to test how different set of features affect the performance of an ML model.

Combination name	Features included
structural characteristics (STRS)	<i>adamic_adar</i> , <i>common_neighbors</i> , <i>preferential_attachment</i> , <i>total_neighbors</i>
structural and textual characteristics (ALL)	<i>adamic_adar</i> , <i>common_neighbors</i> , <i>preferential_attachment</i> , <i>total_neighbors</i> , <i>similarity</i>
<i>adamic_adar</i> and authors similarity (AA-SIM)	<i>adamic_adar</i> , <i>similarity</i>
<i>adamic_adar</i> (AA)	<i>adamic_adar</i>

4.3 Evaluation

To evaluate the effectiveness of our approach, we assess how the performance of various binary classifiers is affected by the ‘*similarity*’ feature. The list of the binary classifiers considered in this paper includes: logistic regression (LR), k-nearest neighbors (50NN), linear support vector machines (LSVM), decision tree (DT) and neural networks (NN). An extensive list of experiments using various classifiers along with different hyperparameter configurations can be found on the GitHub repository of this paper (https://github.com/NCODER/CORD19_GraphOfDocs). Our performance metrics include the *accuracy*, *precision* and *recall* of the binary classifiers. The Friedman

Table 4. Mean (AVG), minimum (MIN), maximum (MAX) and standard deviation (SD) of accuracy, precision and recall metrics per text classifier for each combination of selected features on the nine different datasets. Bold font indicates the best method for each ML model as far as the mean and the standard deviation value of each individual metric are concerned.

Method	Accuracy				Precision				Recall				
	AVG	MIN	MAX	SD	AVG	MIN	MAX	SD	AVG	MIN	MAX	SD	
LR	STRS	0.955	0.938	0.965	0.007	0.955	0.907	0.981	0.022	0.955	0.933	0.976	0.013
	ALL	0.959	0.942	0.968	0.007	0.961	0.909	0.984	0.023	0.957	0.940	0.982	0.013
	AA-SIM	0.972	0.964	0.981	0.005	0.977	0.962	0.991	0.011	0.969	0.955	0.981	0.007
50NN	AA	0.971	0.964	0.978	0.005	0.968	0.945	0.985	0.014	0.974	0.964	0.986	0.008
	STRS	0.956	0.925	0.975	0.014	0.928	0.872	0.965	0.025	0.988	0.982	0.997	0.005
	ALL	0.959	0.925	0.976	0.015	0.933	0.872	0.966	0.026	0.989	0.982	0.996	0.004
LSVM	AA-SIM	0.967	0.948	0.979	0.010	0.944	0.908	0.969	0.018	0.993	0.990	0.996	0.002
	AA	0.957	0.941	0.969	0.010	0.927	0.895	0.947	0.019	0.992	0.987	0.999	0.004
	STRS	0.959	0.931	0.971	0.012	0.941	0.884	0.972	0.026	0.979	0.969	0.993	0.008
LSVM	ALL	0.963	0.936	0.976	0.012	0.946	0.888	0.975	0.026	0.983	0.973	0.996	0.007
	AA-SIM	0.973	0.957	0.981	0.009	0.957	0.926	0.977	0.017	0.989	0.985	0.994	0.003
	AA	0.968	0.953	0.980	0.009	0.952	0.920	0.975	0.018	0.987	0.981	0.994	0.004

(continued)

Table 4. (continued)

Method	Accuracy					Precision					Recall			
	AVG	MIN	MAX	SD		AVG	MIN	MAX	SD		AVG	MIN	MAX	SD
DT	STRS	0.922	0.826	0.979	0.057	0.878	0.742	0.967	0.084		0.994	0.989	1	0.004
	ALL	0.933	0.837	0.980	0.046	0.891	0.755	0.969	0.070		0.994	0.991	0.999	0.003
	AA-SIM	0.931	0.836	0.972	0.045	0.887	0.754	0.955	0.068		0.995	0.991	0.999	0.002
	AA	0.879	0.660	0.955	0.094	0.825	0.595	0.922	0.108		0.994	0.989	1	0.004
NN	STRS	0.928	0.801	0.982	0.061	0.888	0.715	0.975	0.087		0.993	0.988	0.999	0.003
	ALL	0.938	0.807	0.979	0.054	0.902	0.721	0.971	0.078		0.993	0.988	0.999	0.003
	AA-SIM	0.965	0.943	0.979	0.012	0.941	0.898	0.968	0.022		0.994	0.992	0.999	0.002
	AA	0.956	0.899	0.976	0.024	0.928	0.834	0.968	0.041		0.991	0.985	0.999	0.004

test and the post-hoc test of Nemenyi (alpha value 0.05) are also used to calculate the significant importance between the evaluated approaches.

The obtained results indicate that the inclusion of the ‘similarity’ feature (i) increases the average accuracy, precision and recall scores, and (ii) decreases the standard deviation of the aforementioned scores (Table 4). The decrement of the standard deviation in the accuracy score indicates that our approach is reliable regardless of the size of the given dataset. Furthermore, by comparing the average precision score to the average recall score, we conclude that our approach predicts most of the future collaborations correctly. The best average accuracy score is achieved by the LSVM classifier, using the ‘adamic_adar’ and the ‘similarity’ features. Hence, the combination of these two features seems to be the most appropriate one. On the contrary, features such as ‘common_neighbors’, ‘preferential_attachment’ and ‘total_neighbors’ add noise to the overall link prediction process.

Our approach differs from existing ones in that it considers both the textual similarity between the abstracts of the papers for each pair of authors and the structural characteristics of the associated ‘Author’ nodes, aiming to predict a future collaboration between them. The utilization of the textual information in combination with the structural information of a scientific knowledge graph results in better and more reliable ML models, which are less prone to overfitting. Contrary to existing algorithms for the discovery of future research collaborations, our approach exploits structural characteristics and does not ignore the importance of the information related to the unstructured text of papers written by authors. Finally, existing approaches that concentrate only on the exploitation of unstructured textual data rely heavily on NLP techniques and textual representations, which in turn necessitate the generation of sparse feature spaces; hence, in such approaches, the effects of the ‘curse-of-dimensionality’ phenomenon re-emerge.

5 Conclusions

This paper considers the problem of discovering future research collaborations as a link prediction problem applied on scientific knowledge graphs. The proposed approach integrates into a single knowledge graph both structured and unstructured textual data using the graph-of-docs text representation. For the required experimentations, we generated nine different datasets using the CORD-19 dataset. For evaluation purposes, we benchmarked our approach against several link prediction settings, which use various combinations of a set of available features. The evaluation results demonstrated (i) an improvement of the average accuracy, precision and recall of the future collaborations prediction task, and (ii) a mitigation of the effects of the ‘curse-of-dimensionality’ phenomenon.

In any case, our approach has a performance issue, since the time required to build the scientific knowledge graph increases exponentially with the number of graph nodes. Aiming to address the above limitation, while also enhancing the performance and advancing the applicability of our approach, our future work directions include: (i) the utilization of in-memory graph databases in combination with Neo4j; (ii) the experimentation with word, node and graph embeddings (Mikolov et al. 2013; Nikolentzos et al. 2017; Hamilton et al. 2017); (iii) the integration of other scientific research graphs such

as OpenAIRE (Manghi et al. 2019) and Microsoft Academic Graph (Arnab et al. 2015), and (iv) the integration and meaningful exploitation of our approach into collaborative research environments (Kanterakis et al. 2019).

Acknowledgments. The work presented in this paper is supported by the OpenBio-C project (www.openbio.eu), which is co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (Project id: T1EDK- 05275). The authors would also like to thank Stamatis Karlos for his assistance with the statistical analysis of the data.

References

- Adamic, L.A., Adar, E.: Friends and neighbors on the Web. *Soc. Networks* **25**, 211–230 (2003)
- Aggarwal, C.C.: *Machine Learning for Text*. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-73531-3>
- Albert, R., Barabási, A.: *Statistical mechanics of complex networks*. ArXiv, cond-mat/0106096 (2001)
- Arnab, S., Zhihong, S., Yang Song, H.M., Darrin Eide, B.H., Kuansan, W.: An overview of microsoft academic service (MAS) and applications. In: *Proceedings of the 24th International Conference on World Wide Web (WWW 2015 Companion)*, pp. 243–246. ACM, New York (2015)
- Fire, M., et al.: Link prediction in social networks using computationally efficient topological features. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 73–80 (2011)
- Giarelis, N., Kanakaris, N., Karacapilidis, N.: An innovative graph-based approach to advance feature selection from multiple textual documents. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds.) *AIAI 2020. IAICT*, vol. 583, pp. 96–106. Springer, Cham (2020a). https://doi.org/10.1007/978-3-030-49161-1_9
- Giarelis, N., Kanakaris, N., Karacapilidis, N.: On a novel representation of multiple textual documents in a single graph. In: Czarnowski, I., Howlett, Robert J., Jain, Lakhmi C. (eds.) *IDT 2020. SIST*, vol. 193, pp. 105–115. Springer, Singapore (2020b). https://doi.org/10.1007/978-981-15-5925-9_9
- Guns, R., Rousseau, R.: Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics* **101**(2), 1461–1473 (2014). <https://doi.org/10.1007/s11192-013-1228-9>
- Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034 (2017)
- Huang, J., Zhuang, Z., Li, J., and Giles, C. L.: Collaboration over time: characterizing and modeling network evolution. In: *Proceedings of the 2008 international conference on web search and data mining*, pp. 107–116 (2008)
- Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vandoise Sci Nat* **37**, 547–579 (1901)
- Julian, K., Lu, W.: Application of machine learning to link prediction (2016)
- Kanterakis, A., et al.: Towards reproducible bioinformatics: the OpenBio-C scientific workflow environment. In: *Proceedings of the 19th IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, Athens, Greece, pp. 221–226 (2019)
- Li, S., Huang, J., Zhang, Z., Liu, J., Huang, T., Chen, H.: Similarity-based future common neighbors model for link prediction in complex networks. *Sci. Rep.* **8**, 1–11 (2018)

- Liben-Nowell, D., Kleinberg, J.M.: The link-prediction problem for social networks. *J. Am. Soc. Inform. Sci. Technol.* **58**, 1019–1031 (2007)
- Manghi, P., et al.: OpenAIRE Research Graph Dump (Version 1.0.0-beta) [Data set]. Zenodo. (2019). <http://doi.org/10.5281/zenodo.3516918>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NeurIPS)*, pp. 3111–3119 (2013)
- Nathani, D., Chauhan, J., Sharma, C., Kaul, M.: Learning attention-based embeddings for relation prediction in knowledge graphs. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4710–4723 (2019)
- Nikolentzos, G., Meladianos, P., Vazirgiannis, M.: Matching node embeddings for graph similarity. In: *Thirty-First AAAI Conference on Artificial Intelligence* (2017)
- Panagopoulos, G., Tsatsaronis, G., Varlamis, I.: Detecting rising stars in dynamic collaborative networks. *J. Informetrics* **11**, 198–222 (2017)
- Ponomarev, B., Boardman, C.: What is co-authorship? *Scientometrics* **109**(3), 1939–1963 (2016). <https://doi.org/10.1007/s11192-016-2127-7>
- Rousseau, F., Kiagias, E., Vazirgiannis, M.: Text categorization as a graph classification problem. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 1, pp. 1702–1712 (2015)
- Rousseau, F., Vazirgiannis, M.: Graph-of-word and TW-IDF: new approach to ad hoc IR. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 59–68, ACM Press (2013)
- Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C., Han, J.: Co-author relationship prediction in heterogeneous bibliographic networks. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 121–128 IEEE (2011)
- Vahdati, S., Palma, G., Nath, R.J., Lange, C., Auer, S., Vidal, M.-E.: Unveiling scholarly communities over knowledge graphs. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J.C. (eds.) *TPDL 2018. LNCS*, vol. 11057, pp. 103–115. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00066-0_9
- Vathy-Fogarassy, Á., Abonyi, J.: *Graph-based clustering and data visualization algorithms*. Springer, London (2013). <https://doi.org/10.1007/978-1-4471-5158-6>
- Veira, N., Keng, B., Padmanabhan, K., Veneris, A.: Unsupervised embedding enhancements of knowledge graphs using textual associations. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 5218–5225. AAAI Press (2019)
- Wang, L., et al.: *CORD-19: The Covid-19 Open Research Dataset*. arXiv preprint [arXiv:2004.10706](https://arxiv.org/abs/2004.10706) (2020)
- Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.* **29**(12), 2724–2743 (2017)
- Wang, Z., Li, J., Liu, Z., Tang, J.: Text-enhanced representation learning for knowledge graph. In: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4–17 (2016)
- Yu, Q., Long, C., Lv, Y., Shao, H., He, P., Duan, Z.: Predicting co-author relationship in medical co-authorship networks. *PLoS ONE* **9**(7), 101214 (2014)