Journal of Informetrics 15 (2021) 101126

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Learning multi-resolution representations of research patterns in bibliographic networks

O-Joun Lee [a], Hyeon-Ju Jeon [b], Jason J. Jung [b,*]

[a] Future IT Innovation Laboratory, Pohang University of Science and Technology 77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do 37673, Republic of Korea
[b] Department of Computer Engineering, Chung-Ang University 84, Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea

ARTICLE INFO

Article history:
Received 30 April 2020
Received in revised form 5 December 2020
Accepted 7 December 2020
Available online 19 January 2021

Keywords:
Bibliographic network embedding
Skewed distribution
Multi-resolution representation learning
Level-wise simplification
Outstanding scholars

ABSTRACT

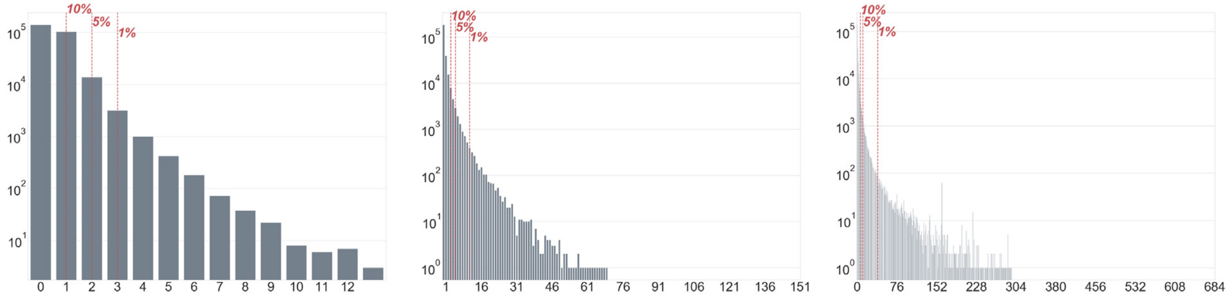
This study aims at representing research patterns of bibliographic entities (e.g., scholars, papers, and venues) with a fixed-length vector. Bibliographic network structures rooted in the entities are incredibly diverse, and this diversity increases in the outstanding entities. Thus, despite their significant volume, the outstanding entities obtain minimal learning opportunities, whereas low-performance entities are over-represented. This study solves the problem by representing the patterns of the entities rather than depicting individual entities in a precise manner. First, we describe structures rooted in the entities using the Weisfeiler–Lehman (WL) relabeling process. Each subgraph generated by the relabeling process provides information on the scholars, kinds of papers they published, standards of venues in which the papers were published, and types of their collaborators. We assume that a subgraph depicts the research patterns of bibliographic entities, such as the preference of a scholar in choosing either a few highly impactful papers or numerous papers of moderate impact. Then, we simplify the subgraphs according to multiple levels of detailedness. Original subgraphs represent the individuality of the entities, and simplified subgraphs represent the entities sharing the same research patterns. In addition, simplified subgraphs balance the learning opportunities of high- and low-performance entities by co-occurring with both types of entities. We embed the subgraphs using the Skip-Gram method. If the results of the embedding represent the research patterns of the entities, the obtained vectors should be able to represent various aspects of the research performance in both the short-term and long-term durations regardless of the performances of the entities. Therefore, we conducted experiments for predicting 23 performance indicators during four time periods for four performance groups (top 1%, 5%, 10%, and all entities) using only the vector representations. The proposed model outperformed the existing network embedding methods in terms of both accuracy and variance.

1. Introduction

With an increase in the number of academic publications, services for discovering capable collaborators (Brand ao, Moro, Lopes, & Oliveira, 2013; Xia, Chen, Wang, Li, & Yang, 2014; Zhou, Ding, Li, & Wan, 2017) and searching of high-quality papers and venues are required (Bai et al., 2019; Cai, Han, & Yang, 2018; Zhang, Zhao, Cheng, Cheng, & Wang, 2016). Such services

[*] Corresponding author.
E-mail addresses: ojlee112358@postech.ac.kr (O.-J. Lee), higd963@gmail.com (H.-J. Jeon), j3ung@cau.ac.kr (J.J. Jung).

https://doi.org/10.1016/j.joi.2020.101126
1751-1577/© 2020 Elsevier Ltd. All rights reserved.



(a) h-index of scholars

(b) Number of papers published by scholars

(c) Number of citations of papers

Fig. 1. Performance distributions of scholars and papers in our dataset (2011–2015). The X- and Y-axes of the plots indicate the indicator values and log-scaled frequency, respectively. The three red-dashed lines indicate boundaries of the top 10%, 5%, and 1% bibliographic entities, respectively, in terms of each performance indicator.

require reliable performance indicators for bibliographic entities (e.g., scholars, papers, and venues). However, existing indicators (Cai, Tian, et al., 2018) only reflect partial aspects of the performance (Jeon, Lee, & Jung, 2019). For example, consider two scholars, a_i and a_j, who have the same h-index. Let us assume that a_i published a few papers that were frequently cited, and a_j published numerous papers that received a relatively lower number of citations. In this case, it would be imperative to know if the level of performance of the two scholars is the same, or if the performance of one scholar is better than the other. Therefore, in our previous study (Jeon et al., 2019), we concentrated on a research performance consisting of various aspects. We also focused on the fact that the scholars have diverse research styles, and that this diversity is shown more clearly by outstanding scholars.

In our previous study (Jeon et al., 2019), we attempted to represent the research styles of different scholars by embedding their co-authorship networks. Conventional indicators (Bordons, Aparicio, González-Albo, & Díaz-Faes, 2015; Reyes-Gonzalez, Gonzalez-Brambila, & Veloso, 2016) are mostly focused on discovering outstanding scholars or research groups. However, using a subgraph-based graph-embedding model, we can analyze the team formation capability of the research groups and the role of the scholars in such groups. By modifying the Weisfeiler–Lehman (WL) relabeling process (Shervashidze, Schweitzer, van Leeuwen, Mehlhorn, & Borgwardt, 2011), we proposed a subgraph extraction method that can cover the size, structure, and cohesion of the research groups. We call such subgraphs 'collaboration patterns.'

By embedding the collaboration patterns, we attempted to discover the research styles of different scholars. Using the results of embedding, we clustered the scholars and found a few interesting clusters; e.g., scholars having several collaborators and papers but relatively fewer citations, and scholars with fewer high-impact papers but a relatively lower social influence. However, the clusters obtained had a low cohesion, and the properties of their elements had a high variance. This problem was mainly caused by skewed distributions of the bibliographic entities (Fig. 1).

Bibliographic networks include numerous entities (e.g., 5,033,294 papers and 2,487,483 scholars indexed in DBLP[1] as of April 2020), most of which are insignificant. As shown in Fig. 1, most of the scholars participated in the publication of only a single paper. In addition, the performance of the outstanding entities is highly diverse compared to their number. In bibliographic network embedding, this skewed distribution accepts the learning opportunities from a few outstanding bibliographic entities. In our previous study (Jeon et al., 2019), this problem made the vector representations of low-performance scholars unnecessarily detailed and high-performance ones indistinct. Because it is well known that the performances of the scholars follow a Pareto distribution (Abramo, Cicero, & D'Angelo, 2013), the skewed distribution may hinder the existing bibliographic network embedding methods and their applications (e.g., author name disambiguation Ma, Wang, & Zhang, 2019; Zhang & Hasan, 2017 in paper recommendations Cai, Han, et al., 2018).

To resolve these limitations, we propose a model for embedding bibliographic networks at multiple resolutions. First, we extend the 'collaboration patterns' to 'research patterns.' This study deals with heterogeneous bibliographic networks that include three types of entities (i.e., scholars, papers, and venues) and two types of relations (i.e., scholar-paper and paper-venue). Thus, the research pattern provides abundant information on the scholars, the papers they published, the venues in which the papers were published, and their collaborators. This information is categorical because the WL relabeling describes subgraphs rooted in each entity using the subgraphs of adjacent entities. The collaborators and venues of the scholars will change frequently, whereas the types of venues, papers, and co-authors preferred by the scholars are relatively consistent. Therefore, we call the subgraphs 'research patterns,' and patterns observed from multiple viewpoints will show the scholars' specific styles. The definitions of the research patterns and methods for extracting them are clarified in Section 3.1.

Subgraphs discovered using WL relabeling can represent network structures in a particular range in a complete and accurate manner. Nevertheless, such accuracy also causes a side effect of the research patterns of outstanding entities not

[1] https://dblp.uni-trier.de/.

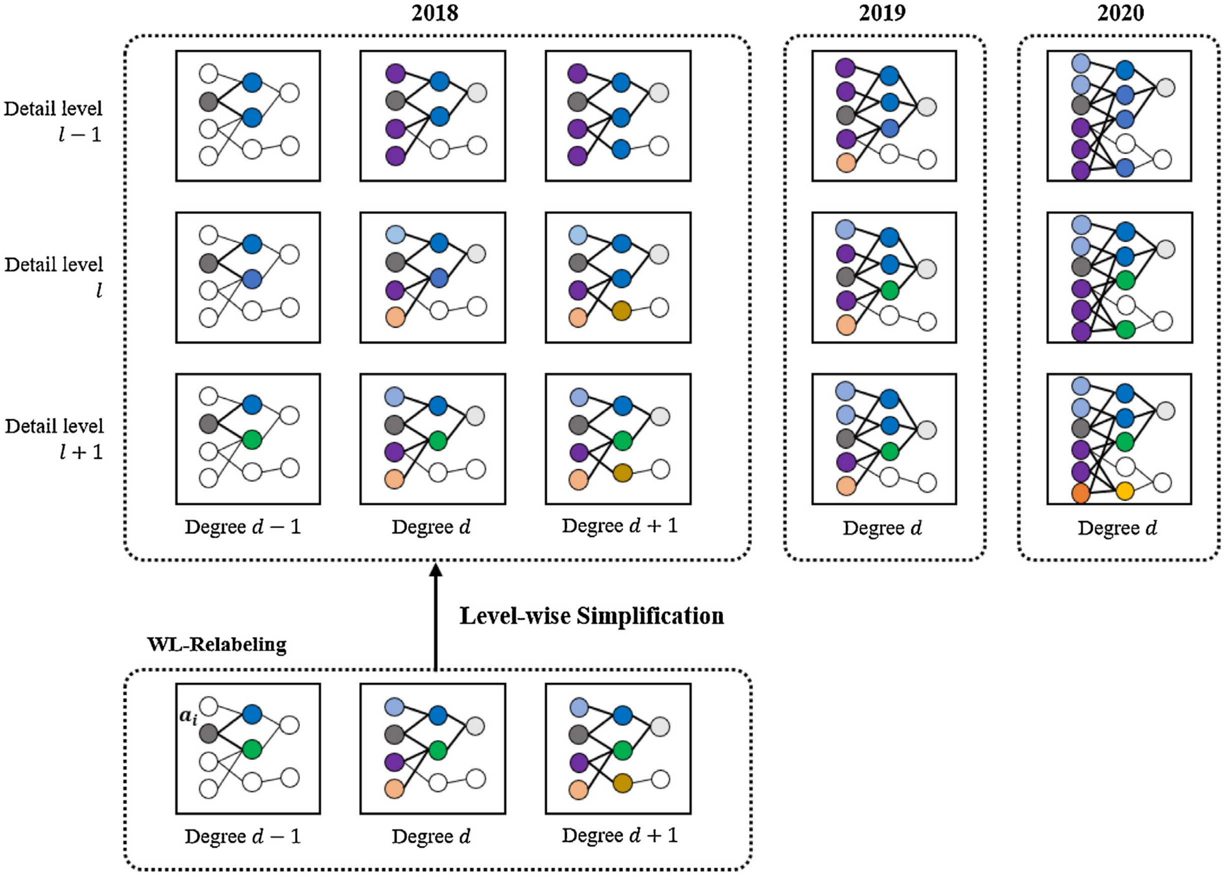


Fig. 2. Conceptual overview of the level-wise simplification. Rectangular boxes indicate research patterns of a scholar a_i on their respective detail levels and degrees. Detail levels indicate the resolution of the research pattern, and the degree indicates the scale of the pattern. The methods for extracting the subgraphs (research patterns) at multiple resolutions and scales are described in Sections 3.2 and 3.1, respectively. Dark gray nodes denote a_i in a bibliographic network. Research patterns of a_i are described based on their neighborhoods (i.e., research patterns rooted in adjacent entities of a_i). By merging similar neighborhoods, the descriptions of the research patterns are simplified (nodes with the same color indicate the same research patterns). At higher detail levels, the research patterns are detailed and diverse. However, at lower detail levels, their research patterns become similar to the reduced diversity of the neighborhoods. In addition, a bibliographic network has temporal dynamics. Research patterns in the higher detail levels will differ for years. However, through simplification, we obtain the research styles of the bibliographic entities, which are consistent.

obtaining sufficient learning opportunities. We have only a few samples for each pattern and cannot conduct oversampling as the bibliographic entities are intrinsic. To solve this problem, we propose a method for a level-wise simplification of the research patterns. This method aims to generate multiple versions of the research patterns by adjusting their detail levels, as shown in Fig. 2. Although there are distinctive patterns on high detail levels, they can be the same after simplification. Thus, simpler research patterns create a connectivity between highly detailed research patterns. Thus, we can balance the learning opportunities of research patterns by embedding the patterns using their co-occurrences.

In addition, structures of bibliographic networks change according to time. Subgraphs rooted in bibliographic entities will be entirely different from the subgraphs present a few years ago, as will their representations. This point might be much more severe for scholars in their early years than in other scholars. If we can determine their styles, these features will be far more constant than the mere network substructures or adjacency. We can discover the styles by removing overabundant information, as shown in Fig. 2. This point will be validated by predicting the future research performances of the entities.

Finally, we can see the bibliographic entities from various viewpoints. Small-scale and simple research patterns prevent the oversimplification of outstanding entities and over-specifying the other entities. Large-scale and complicated patterns preserve details for discovering the research styles of the entities. Through different applications, we can categorize entities with a particular style using multiple resolutions (by decreasing the resolution of entities with uninteresting styles). For example, searching, recommending, and visualizing bibliographic entities are significant for finding academic publications, consulting with scholars, and discovering venues. When we look for collaborators or attempt to recruit a coworker, authorities in their research fields are not always our targets. The similarity in the working styles between an individual and our group is a better paramenter to consider while determining the success of a collaboration or team formation. Therefore, the proposed embedding model can be applied to various applications, from a team formation application (Yu, Bedru, Lee, & Xia,

2019) to the author name disambiguation (Zhang & Hasan, 2017), wherein we need to compare working styles and search for a particular style. This approach is more reasonable than using straightforward weighting factors.

To evaluate the proposed methods, we must validate the following assumptions:

- RQ 1. The research patterns reflect the styles of the bibliographic entities.
- RQ 2. The embedding model is robust to the skewed distribution in the performance of the bibliographic entities.

Because it is difficult to evaluate the accuracy of the vector representations and validate the assumptions using only such representations, we designed an experiment to examine the compliance of the representations with the following requirements:

- The effectiveness of research pattern vectors in predicting various performance indicators (RQ 1).
- The efficiency of research pattern vectors in both long- and short-term predictions of the indicators (RQ 1).
- The use of research pattern vectors in performance prediction of both high- and low-performance bibliographic entities (RQ 2).

There have been various performance indicators for bibliographic entities, and they measure performance through individual viewpoints (Section 2.1). Thus, if the research pattern vectors reflect the research styles, the vectors will correlate with various types of performance indicators. In addition, since the research styles will have a higher consistency in the performance at each moment, the research pattern vectors will be useful for predicting the performance in the future. Finally, if the model is robust, the vector representations should be able to distinguish the styles of the entities regardless of their performance. The robustness indicates that the level-wise simplification effectively handles the skewed and imbalanced distribution of bibliographic entities.

Therefore, we measured 23 indicators for the bibliographic entities for four time periods: 2011–2015, 2011–2016, 2011–2017, and 2011–2018. We also composed four groups of entities according to their performance indicators (top 1%, top 5%, top 10%, and all entities). Research pattern vectors were trained using data only from 2011 to 2015. We then examined whether the research pattern vectors can show a high accuracy on all cases consistently. Section 4.2 presents the experimental procedures in detail.

The remainder of this paper is organized as follows. Section 2 presents existing studies related to the quantitative indicators and bibliographic network embedding. In Section 3, we propose a method for discovering research patterns at various scales and simplifying the patterns according to the detail levels. In addition, we propose a model for learning the multi-resolution representation of research patterns. In Section 4, we verify the effectiveness of the proposed methods using a real bibliographic network. Finally, Section 5 presents some concluding remarks and further research directions of this study.

2. Related studies

2.1. Quantifying the academic impact based on a performance measurement

Research results are distributed through various types of academic publications. The impact of such publications is also the impact of their authors and venues. Measuring the impact is necessary for providing adequate bibliographic entities (i.e., academic publications, their authors, and their venues) to scholars (Bai et al., 2019; Cai, Tian, et al., 2018). Conventional measurements (Cai, Tian, et al., 2018; Waltman, 2016) focus on how much a publication influences other publications. Thus, the number of citations is the most fundamental method for measuring the performance of every type of bibliographic entity.

To evaluate the scholars and venues, the number of papers published is also as significant as the number of citations. By combining the information, we can determine the average quality of their papers. The *h*-index (Hirsch, 2005) and *gh*-index (Galam, 2011) are indicators for assigning high scores to the scholars with high-impact paper publications. However, such methods give the same scores to both the promising young scholars and low-performance scholars. Thus, a few studies (Abbasi, Altmann, & Hwang, 2009; Lippi & Mattiuzzi, 2017; Vaidya, 2005) have attempted to solve this problem by considering the lengths of their academic careers. For the venues, the journal impact factor (Garfield, 2006) is the most well-known indicator. This metric is based on the average number of citations for papers published in each venue. However, the number of citations is affected by traditions and customs specific to the research fields. If most papers in a field have relatively fewer references, the papers will be cited infrequently on an average. Field-weighted indicators (Ferrer-Sapena & Pérez, 2019) have been proposed to solve this problem for journal measurement. For paper-level indicators, Scopus[2] has been providing FWCI (Field-Weighted Citation Impact) (BV, 2018) to compare papers regardless of the research field, publication type, publication age, or other factors.

The above mentioned indicators can be calculated using simple aggregation operations. However, the various aspects of research make it difficult to represent the performance using a single value. For example, citations in papers have different

[2] https://www.scopus.com/.

significance and meaning from each other. Although aggregation approaches make no distinction between them, we can partially examine their differences through the citation network structures. Bergstrom, West, and Wiseman (2008) applied weighting factors to each citation. They supposed that citations in frequently cited papers are more valuable than those in infrequently cited papers. ANI (Article Network Influence) (Chang, Phoa, & Nakano, 2019) applies affinity propagation to a citation network. Similar attempts have also been made in a co-authorship network (Perianes-Rodríguez, Chinchilla-Rodríguez, Vargas-Quesada, Gómez, & Moya-Anegón, 2009; Yan, Zhai, & Fan, 2013).

There have also been studies on the combination of two approaches. The *PR*-index (Gao, Wang, Li, Zhang, & Zeng, 2016) is a combination of an *h*-index and the PageRank centrality measured in a citation network between scholars. A few studies (Tang, Jin, & Zhang, 2008; Zhang, Ma, Wang, Chen, & Yu, 2017) have attempted to consider the research topics of the papers and scholars for their assessment. By applying topic modeling methods (e.g., Latent Dirichlet Allocation) to different papers, these studies obtained the topic distributions of the papers and scholars. The topic distributions were then used as the weighting factors for the PageRank centrality on the citation network.

Quantitative indicators are based on various data sources and are used to analyze the data through individual perspectives regarding the research performance. Therefore, it is difficult to state the effectiveness of each perspective compared to the rest. These heuristic and hand-crafted indicators should be re-designed according to the consensus of the academic community regarding the research performance. If we can represent the various aspects of the research performance and capacity with a fixed-length vector, we will be able to reduce the efforts for a reformation.

2.2. *Bibliographic network embedding*

Various embedding models for general networks have been applied to a bibliographic network. LINE (Tang et al., 2015) and SDNE (Wang, Cui, & Zhu, 2016) are popular methods that embed nodes based on their first- and second-order proximities. These methods aim to embed social networks and assume that the *N*th-order proximity reflects the network structures. Nevertheless, the structures cannot be explicitly considered. To solve this problem, DeepWalk (Perozzi, Al-Rfou, & Skiena, 2014) describes network structures around nodes using the random walk approach. This method also interprets the representations of the nodes using Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and applies inclusive relations between random walks and nodes. However, random walks cannot describe nodes using consistent criteria. Node2Vec (Grover & Leskovec, 2016) and Struc2Vec (Ribeiro, Saverese, & Figueiredo, 2017) developed improved methods for generating random walks. Moreover, Metapath2Vec (Dong, Chawla, & Swami, 2017) and HIN2Vec (yang Fu, Lee, & Lei, 2017) employed metapaths to apply random walk approaches to heterogeneous networks. In contrast, Deep Graph Kernels (Yanardag & Vishwanathan, 2015) and Subgraph2Vec (Narayanan, Chandramohan, Chen, Liu, & Saminathan, 2016) have attempted to describe explicit network structures using consistent criteria. In particular, Subgraph2Vec used the WL relabeling process (Shervashidze et al., 2011) to discover subgraphs rooted in each node at various scales. This approach can be modified according to the type of networks. Therefore, a few studies on extending the WL relabeling process to cover both the node proximity and the temporal dynamics of the network structures have been conducted (Jeon et al., 2019; Lee & Jung, 2020a; Lee, Jung, & Kim, 2020). Based on the WL relabeling, Du and Tong (2019) proposed a method for embedding multiple networks at multiple resolutions. Their motivation is similar to ours, i.e., representing entities present in networks through various resolutions ranging from coarser to finer ones. However, they implemented multiple resolutions by defining entities at various granularity levels (e.g., nodes, subgraphs, and networks). On the other hand, we adjust the scales and detail levels of the subgraphs. This approach is difficult to apply to bibliographic networks owing to several problems. Entities in bibliographic networks are unique, and most of the entities might be connected. Thus, it is challenging to presume multiple and independent bibliographic networks. In addition, this method cannot deal with the skewed distribution of bibliographic entities or the excessive diversity of outstanding entities.

There have also been embedding models specialized for bibliographic networks. Several studies have attempted to reflect not only the network structure but also texts in bibliographic entities. Author2vec (Ganguly, Gupta, Varma, & Pudi, 2016) interprets the representations of scholars by analyzing co-authorship networks and the titles of papers authored by such scholars. This method extracts the topical similarity between papers by using the terms present in the titles. Paper2vec (Ganguly & Pudi, 2017) applies the same strategy for citation networks. However, both these methods can handle only homogeneous bibliographic networks. Zhang, Zhao, and Lu (2019) attempted to capture the content of the papers by integrating the terms in the papers into citation networks. Their model learns three types of relations, i.e., (i) co-occurrences between terms, (ii) the occurrence of terms in different papers, and (iii) citations between papers. GanHBNR (Cai, Han, et al., 2018) and BNR (Cai, Zheng, Yang, Dai, & Guo, 2019) are embedding models used for citation recommendation. They embed bibliographic entities using three types of relations: (i) intra-vertex relations (e.g., paper-to-paper), (ii) inter-vertex relations (e.g., paper-to-author), and (iii) vertex-content relations (e.g., papers-to-terms).

By employing the terms in different papers, several issues arise. For example, the detection of terminologies and the handling of neologisms are difficult to achieve. Therefore, various studies have assumed that the structures of bibliographic entities imply correlations in the content of bibliographic entities. If two papers have similar author lists and the same venue, they will have similar content. Li, Zhang, Yu, Zhang, and Yan (2018) proposed a dynamic network embedding model for human contact networks, co-authorship networks, and e-mail networks. They also conducted link prediction using their embedding model. Jamil, Khan, Halim, and Baig (2011) extracted subgraphs from co-authorship networks by using frequent pattern mining techniques. Based on the subgraphs, they also proposed a network embedding method for predicting co-authorship.

Xiao et al. (2019) combined representations of the scholars, papers, and research topics to predict the performance of the scholars and the quality of the papers. The combination method focused on the influence of scholars on the papers using the co-ranking method (Zhou, Orshanskiy, Zha, & Giles, 2007). Although most of the bibliographic network embedding models are unsupervised, Chen and Sun (2017) proposed a task-specific embedding model for author identification. Their bibliographic network consists of the authors, keywords, publication years, and venues of the papers. This model interprets the research fields of scholars using authorship relations and global network structures using metapaths.

Bibliographic network embedding can capture the styles of bibliographic entities, such as the active collaboration of a scholar with other scholars or the topical diversity of the venue. However, these models do not aim to discover the styles of bibliographic entities themselves. Our previous studies (Jeon et al., 2019) focused on the existing performance indicators that provide minimal consideration to the various aspects of research. There are various types of research fields, scholars, and academic papers. If we merely count the numbers of papers and citations, the results will be significantly biased. We attempted to embed scholars by only using the information on ‘a scholar published which types of papers at which types of venues and with what kinds of collaborators.’ From the embedding vectors, we discovered a few dozen research styles of scholars and indicated that the styles correlate with the performance of the scholars. This study expands on the concept of research styles of scholars to include every kind of bibliographic entity. In addition, as shown in Fig. 1, the skewed distribution of bibliographic entities disturbed the clustering of scholars according to their research styles. This problem is particularly severe in outstanding entities that have too small a volume and too high a diversity.

Most conventional approaches for handling skewed and imbalanced datasets have applied data augmentation or oversampling for scarce entities (Anil & Singh, 2020; Huang, Li, Loy, & Tang, 2016; Khan, Hayat, Bennamoun, Sohel, & Togneri, 2018). Nevertheless, entities in bibliographic networks (e.g., scholars, papers, and venues) are unique. Thus, it is difficult to decide which entities are scarce. If we apply conventional network embedding methods based on an N-th order proximity (Tang et al., 2015; Wang et al., 2016), this problem will not be solved regardless of how large N is. With this approach, at least, scarce entities will have representations which are distinct from the others. However, for the embedding of bibliographic networks, network structures are more significant than the node proximity. Even if a biologist has frequently co-worked with a computer scientist on bioinformatics studies, we cannot state they are similar researchers. From this perspective, we use substructure-based network embedding and reduce the scarcity by simplifying the substructures.

3. Multi-resolution representations of research patterns

Difficulties in analyzing bibliographic networks are mainly derived from (i) heterogeneity and (ii) skewed distributions of bibliographic entities. The first issue has been resolved through the advent of various heterogeneous network embedding models (Cai, Han, et al., 2018; Chen & Sun, 2017). However, to the best of our knowledge, there have not been many studies dealing with the second issue.

As shown in Fig. 1, the performance distributions of scholars and papers are highly skewed in most of their statistical features. Most scholars have contributed in only a single paper, and most papers are never cited. By contrast, outstanding scholars and papers are scarce and diverse. This problem influences the learning representations of bibliographic entities (e.g., scholars, papers, venues, co-authorship, and citations). Low-performance entities obtain several opportunities for learning because they are less distinctive (i.e., they have numerous similar entities) and appear more frequently than the high-performance entities. In contrast, outstanding entities cannot obtain sufficient learning opportunities owing to their scarcity and volume.

We solved this problem by adopting a level-wise simplification. This approach enables us to (i) increase the frequency of occurrence of the scarce entities and (ii) reduce the level of attention for unnecessarily frequent entities. We call these simplification steps ‘detail levels.’ We thus propose a model for the learning of multi-resolution representations of the entities, which reflect the structure of bibliographic networks at various detail levels. The proposed model aims to preserve the details of the structures while balancing the learning opportunities. Fig. 3 presents components of the proposed model and relations between the components.

3.1. Discovering research patterns

In our previous study (Jeon et al., 2019), we extracted subgraphs (collaboration patterns) from a co-authorship network. The collaboration patterns covered only (i) team formations of research groups and (ii) overlaps between groups of scholars. This study extends the collaboration patterns to research patterns by incorporating information regarding papers and venues. When two papers published by the same scholars have disparate venues, the collaboration patterns only indicate that the scholars collaborated twice. However, research patterns will show that the scholars might have interests in two different topics. In other words, the research patterns give information on types of the scholars, kinds of the papers they published, standards of the venues in which the papers were published, and types of their collaborators. Our bibliographic network is similar to a tripartite graph as follows:

Definition 1 *(Bibliographic Network).* The bibliographic network ($\mathcal{N}$) contains three types of nodes: scholars ($\mathbb{A}$), papers ($\mathbb{P}$), and venues ($\mathbb{V}$). Between these nodes, there are two types of relations: An author ‘writes’ a paper ($\mathcal{W} \in \mathbb{R}^{|\mathbb{A}|\times|\mathbb{P}|}$), and a

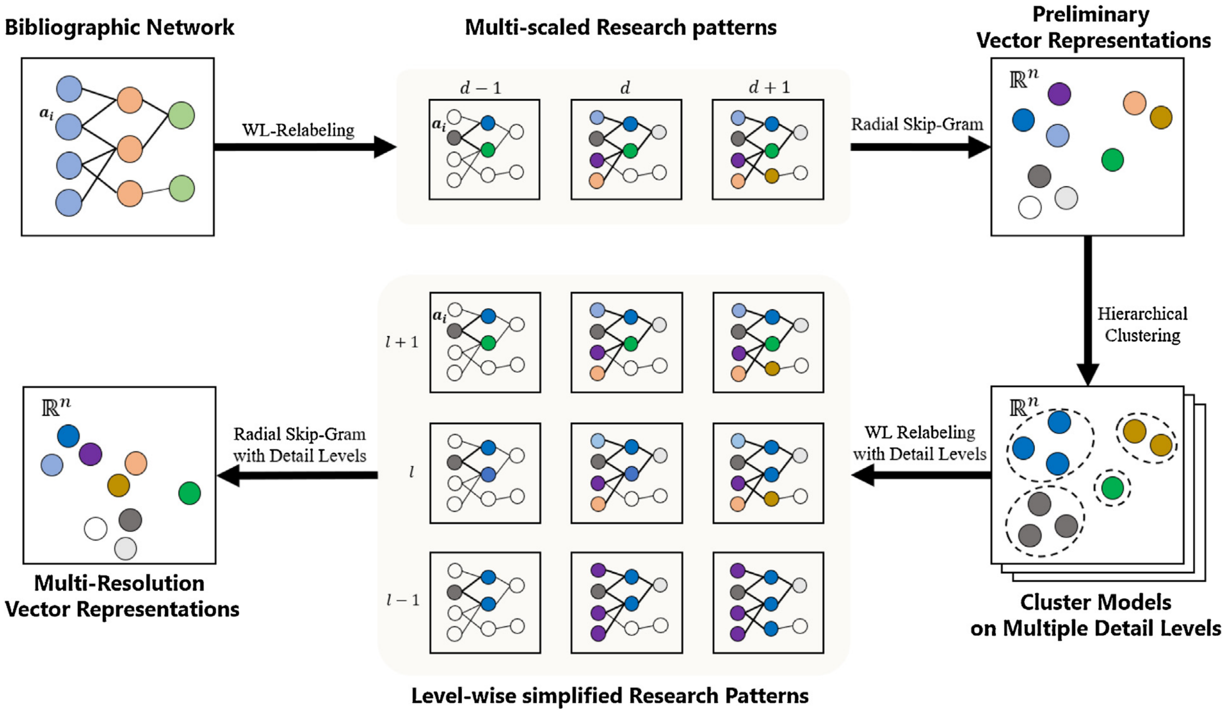
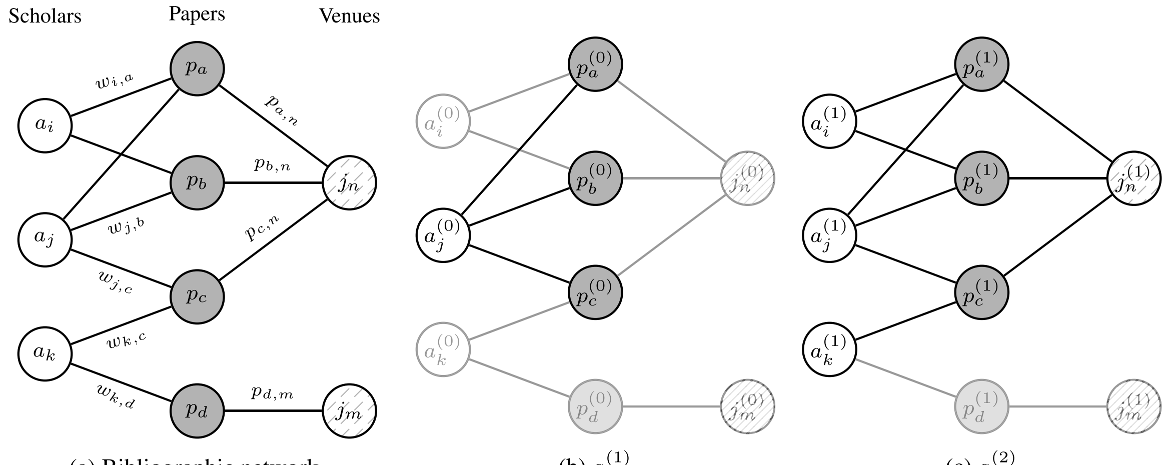

Bibliographic Network
a_i
WL-Relabeling
Multi-scaled Research patterns
$d-1$
d
$d+1$
Radial Skip-Gram
Preliminary Vector Representations
$\mathbb{R}^n$
Hierarchical Clustering
Cluster Models on Multiple Detail Levels
WL Relabeling with Detail Levels
$l+1$
l
$l-1$
Level-wise simplified Research Patterns
Radial Skip-Gram with Detail Levels
Multi-Resolution Vector Representations

Fig. 3. Overview of the proposed model. Our bibliographic model is defined in Section 3.1, and this section also describes methods for extracting and embedding research patterns rooted in each bibliographic entity to acquire their preliminary vector representations. Section 3.2 presents a method for level-wise simplification of research patterns by applying the hierarchical clustering to the preliminary vector representations. Finally, in Section 3.3, we describe a method for learning multi-resolution representations of bibliographic entities.



(a) Bibliographic network (b) $a_j^{(1)}$ (c) $a_j^{(2)}$

Fig. 4. Example of a bibliographic network and research patterns of scholars: (a)–(c) represent changes in research patterns according to degrees ranging from 0 to 2. White nodes indicate the scholars, gray nodes denote the papers, and shaded nodes indicate the venues.

paper is published in a venue ($\mathcal{P} \in \mathbb{R}^{|\mathbb{P}|\times|\mathbb{V}|}$). The two relations have weights ($\in \{0, 1\}$) that indicate only the existence of the relations. This can be formulated as follows:

$$\mathcal{N} = \langle \mathbb{A}, \mathbb{P}, \mathbb{V}, \mathcal{W}, \mathcal{P} \rangle. \tag{1}$$

Fig. 4(a) illustrates an example of a bibliographic network, where $a_i \in \mathbb{A}$, $p_a \in \mathbb{P}$, $v_n \in \mathbb{V}$, $w_{a,i} \in \mathcal{W}$, and $p_{a,n} \in \mathcal{P}$.

Research patterns are subgraphs rooted in entities in a bibliographic network. The patterns indicate the performances of bibliographic entities as the outstanding scholars may publish impactful papers at high-quality venues. However, if one scholar writes numerous moderate-quality papers, and another scholar writes a few high-quality papers, both will exhibit high-performance. They have quite different research styles and research patterns despite having a similar research outcome. Therefore, the research patterns also reflect different styles of the entities.

To extract the research patterns, we modify the WL relabeling process, which is based on the WL graph isomorphism test (Shervashidze et al., 2011). This method is used to describe structures around entities at multiple scales in both a strict and precise manner. However, when the structures are highly detailed, the WL relabeling would be incapable of adjusting the detail levels of the structures. Therefore, we first compose research patterns using the WL relabeling process. We then simplify the patterns according to the various detail levels (Section 3.2). At lower detail levels, the research patterns are more simplified and less diverse. However, the patterns at higher detail levels are closer to the original patterns. At the highest detail level, the research pattern is defined as follows:

Definition 2 *(Research Pattern).* Consider $a_i^{(d)}$ indicates the research pattern of a_i for degree $d \in [0, D]$. The research patterns rooted in a_i represent (i) collaborators a_i, (ii) papers written by a_i, and (iii) the venues publishing the papers. The degree provides (iv) coverage on the research patterns, which are observation points for discovering the patterns. A research pattern for degree d is described based on (i) itself and (ii) its neighborhoods for degrees $d-1$. When $w_{i,a}$ indicates a relation in which author a_i writes a paper p_a, and $\mathcal{S}_i^{(d-1)}$ denotes a set of subgraphs rooted in all possible p_a for degree $d-1$, $a_i^{(d)}$ can be formulated as follows:

$$a_i^{(d)} = \langle a_i^{(d-1)}, \mathcal{S}_i^{(d-1)} \rangle, \mathcal{S}_i^{(d-1)} = \left\{ p_a^{(d-1)} \mid w_{i,a} \neq 0, \forall p_a \right\}, \tag{2}$$

where $p_a^{(d-1)}$ refers to a research pattern of p_a for degrees $d-1$.

Definition 2 describes the research pattern using only an example of the scholars. The same method is also applied to the other types of bibliographic entities (e.g., papers and venues). The only difference is that the papers have two types of neighborhoods (both scholars and venues). For a paper p_a, $\mathcal{S}_a^{(d-1)} = \{a_i^{(d-1)} \mid w_{i,a} \neq 0, \forall a_i\} \cup \{j_n^{(d-1)} \mid p_{a,n} \neq 0, \forall j_n\}$. For a degree of 0, all entities are labeled based on their types (e.g., $\mathbb{A}$, $\mathbb{P}$, and $\mathbb{V}$). As the degree increases, subgraphs represent broader and more subdivided local structures around the entities. Fig. 4(b) and (c) illustrates how a one-hop connectivity at degree 1 induces a two-hop connectivity at degree 2. The WL relabeling process is conducted in an iterative manner. We assign labels for all subgraphs at a degree of d based on the subgraphs for degrees of $d-1$ at the d-th iteration. Although these labels are nominal, they are not unique for each entity. To manage the labels, we compose a dictionary ($\mathcal{S}$) that has descriptions of the subgraphs and a hash of the descriptions as values and keys, e.g., $\{HASH(a_i^{(d)}) : a_i^{(d)}\}$. Algorithm 3 shows all the procedures used for discovering the research patterns. Furthermore, we conducted a parameter search for determining the maximum degree D (Section 4.1).

Algorithm 3. WL relabeling process for discovering research patterns.

```
procedure WLreLABELLING$\mathcal{N}, \mathcal{S}$
    for $d : 1 \rightarrow D$ do
        for $a_i \in \mathbb{A}$ do
            $\mathcal{S}_i^{(d-1)} \longleftarrow \{p_a^{(d-1)} \mid w_{i,a} \neq 0, \forall p_a\}$
            $a_i^{(d)} \longleftarrow \langle a_i^{(d-1)}, \mathcal{S}_i^{(d-1)} \rangle$
            Put $\left\{ HASH\left(a_i^{(d)}\right) : a_i^{(d)} \right\}$ into $\mathcal{S}$
        end for
        for $p_a \in \mathbb{P}$ do
            $\mathcal{S}_a^{(d-1)} \longleftarrow \{a_i^{(d-1)} \mid w_{i,a} \neq 0, \forall a_i\} \cup \{j_n^{(d-1)} \mid p_{a,n} \neq 0, \forall j_n\}$
            $p_a^{(d)} \longleftarrow \langle p_a^{(d-1)}, \mathcal{S}_a^{(d-1)} \rangle$
            Put $\left\{ HASH\left(p_a^{(d)}\right) : p_a^{(d)} \right\}$ into $\mathcal{S}$
        end for
        for $j_n \in \mathbb{J}$ do
            $\mathcal{S}_n^{(d-1)} \longleftarrow \{p_a^{(d-1)} \mid p_{a,n} \neq 0, \forall p_a\}$
            $j_n^{(d)} \longleftarrow \langle j_n^{(d-1)}, \mathcal{S}_n^{(d-1)} \rangle$
            Put $\left\{ HASH\left(j_n^{(d)}\right) : j_n^{(d)} \right\}$ into $\mathcal{S}$
        end for
    end for
end procedure
```

3.2. Simplifying research patterns with detail levels

Using the WL relabeling process, each bibliographic entity is represented using multiple sets of research patterns (subgraphs), similar to sentences and words. Since the research patterns are discrete, they should be associated with a sufficient number of bibliographic entities to learn their vector representations. However, at a higher degree, a greater number of diverse research patterns appear in fewer bibliographic entities. In our dataset, duplication ratios of research patterns[3] at degrees of 1 and 4 were 99.96% and 55.10%, respectively. In addition, owing to the skewed distribution (in Fig. 1), the scarcity

[3] A ratio of research patterns that appear in plural bibliographic entities for all research patterns.

of research patterns is much more severe in the high-performance entities than in the low-performance entities. For the top 10% of scholars in terms of the number of paper publications, 6.56% of their research patterns appeared in plural scholars. In contrast, for scholars who wrote only one paper (at nearly 90%), the ratio was 94.99%.

As discussed in the previous sections, this sparsity of duplication is harmful to representational learning. However, abandoning the preciseness of the research patterns cannot be an effective solution. Therefore, we focus on Subgraph2Vec (Narayanan et al., 2016), which is a well-known model for learning representations of subgraphs and providing connectivity between high-degree subgraphs based on low-degree subgraphs. This model learns changes in the subgraphs according to their degrees by defining the adjacency between subgraphs based on (i) their degrees and (ii) the adjacency of the nodes that they are rooted in. In this study, we extend the definition of adjacency by adding one more criterion, i.e., the detail levels. We anticipate that simpler patterns will provide indirect connectivity between finer patterns. In other words, the adjacency of the detail levels will provide learning opportunities to the finer patterns. When two scholars a_i and a_j have research patterns $\langle s_A, s_B, s_C\rangle$ and $\langle s_D, s_E, s_F\rangle$, respectively, the patterns indicate that a_i and a_j are completely different. However, if a_i and a_j are simplified to $\langle s_\alpha, s_\beta, s_\gamma\rangle$ and $\langle s_\alpha, s_\delta, s_\eta\rangle$, respectively, we can find a similarity between a_i and a_j from s_α. In addition, we can have opportunities for learning the distance between s_A and s_D through their co-occurrence for s_α. This opportunity will be propagated to the other research patterns of a_i and a_j as well. When $a_{i,l}^{(d)}$ indicates a simplified research pattern of a_i at a degree d and a detail level l, the neighborhood of the pattern can be formulated as follows:

$$\mathcal{N}_{i,l}^{(d)} = \left\{ p_{a,l+\Delta l}^{(d+\Delta d)} \;\middle|\; w_{i,a} \neq 0, |\Delta l| \leq \mathcal{W}_L, |\Delta d| \leq \mathcal{W}_D, \forall p_a \right\}, \tag{3}$$

where $\mathcal{W}_L$ and $\mathcal{W}_D$ are the window sizes for the detail level and degree, respectively. Neighborhoods of other types of bibliographic entities are composed in the same way.

We propose a method for simplifying the research patterns according to the detail levels. To simplify the substructures (e.g., subgraphs and paths) in bibliographic networks, the most widely-used approach is the conversion of the paths consisting of heterogeneous nodes and relations into new relations (e.g., a co-authorship). As a result, bibliographic networks are transformed into co-authorship, citation, and other types of networks (Zhou et al., 2017). However, this approach is inadequate for gradually adjusting the detail levels. In addition, because this transformation uses predefined meta-paths, it is difficult to deal with diversity and ambiguity in the research styles. The clustering of homogeneous entities is also a widely used approach because bibliographic networks contain numerous insignificant entities and relations (Loudcher, Jakawat, Morales, & Favre, 2015; Wu, Lin, Wang, & Gregory, 2016). This approach is mostly conducted by merging nodes in the same clusters or communities. However, because our model assumes that bibliographic entities have different meanings according to their degrees (observation ranges) and detail levels, it is inadequate to continuously omit the entities on a level-by-level basis.

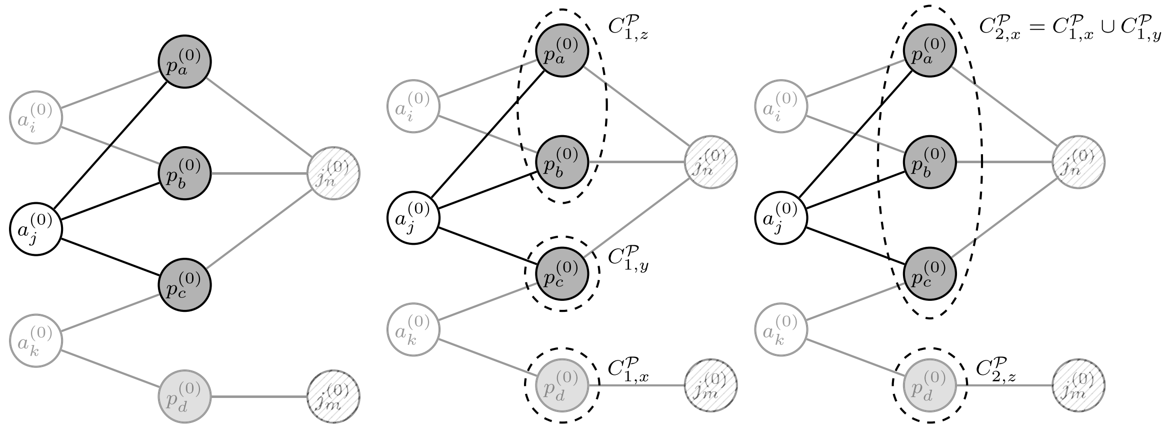
If the description of $a_i^{(d)}$ consists mostly of similar research patterns at a degree of $d-1$, annotating the labels of all patterns at a degree of $d-1$ will provide an overabundance of information. In addition, distinctive neighbors will be more helpful in describing research patterns than numerous ordinary neighbors. Therefore, to simplify the research patterns for a degree of d, we cluster patterns for a degree of $d-1$ with multiple resolutions. We then replace the patterns for a degree of $d-1$, which are in the descriptions for a degree of d, by their clusters. As the clusters decrease in number and become coarser, the descriptions become more simplified and compressed. Thus, we can reduce the diversity of the research patterns without merging the patterns or entities (e.g., scholars, papers, and venues).

For level-wise simplification, the hierarchical clustering will be an effective method. However, the bibliographic networks consist of innumerable nodes and edges; hence, our dataset includes 487,462 nodes. Thus, hierarchical clustering requires an overabundance of computational resources as it has to update the distances between all the clusters during each iteration. Even if we use other clustering methods, there are limitations in searching the optimal number of clusters. To avoid this problem, we merely fixed the number of clusters by interconnecting the detail levels.

In this study, we cluster the research patterns into 2^l clusters according to their similarity using the expectation-maximization (EM) algorithm, in which l is a detail level. EM is less sensitive to the ranges and distributions of the components of the feature vectors than other conventional clustering methods. To measure the similarity, we compose a preliminary version of the vector representation of the research patterns using Subgraph2Vec (Narayanan et al., 2016) (Eq. (5)), which does not consider the detail levels. Clustering is conducted for each degree and at each detail level. By adjusting l, we can obtain the clusters at various resolutions.

In brief, the research patterns are simplified by removing duplicate neighborhoods from their descriptions. For describing $a_i^{(d)}$, if $a_i^{(d-1)}$ have multiple neighborhoods in the same cluster, we do not need to annotate the adjacency multiple times. Fig. 5 illustrates how $a_j^{(1)}$ in Fig. 4 becomes simpler in accordance with the detail levels. We conduct this simplification on every degree and at every detail level. At a lower detail level, the number of clusters decreases, and the descriptions of the research patterns become shorter and simpler. Finally, the neighborhoods will be integrated into a single neighborhood at a detail level of 0; in addition, the research patterns become similar to the node types. For the detail levels, the definition of the research patterns in Definition 2 is extended as follows:

Definition 4 *(Research pattern with detail levels).* Suppose that $a_{i,l}^{(d)}$ indicates a research pattern of a_i for a degree of d at a detail level of $l \in [0, L]$. At a detail level of l, the number of clusters is 2^l. When $2^l \geq |\mathbb{P}|$, $a_{i,l}^{(d)} = a_i^{(d)}$. As l increases in size, the



(a) $a_{j,L+1}^{(1)} = a_j^{(1)}$ (b) $a_{j,L}^{(1)}$ (c) $a_{j,L-1}^{(1)}$

Fig. 5. Example of a level-wise simplification of research patterns rooted for scholars: (a)–(c) represent changes in the research patterns in accordance with the decreases in the detail levels from $L+1$ to $L-1$ through L. At the detail level $L+1$, the research patterns are the same as they were before the simplification process.

structural information depicted by $a_{i,l}^{(d)}$ becomes more detailed. When $C_l^{\mathcal{P}}$ is a cluster model for papers at a detail level of l, $a_{i,l}^{(d)}$ is described based on the intersections between $\mathcal{S}_i^{(d-1)}$ and the clusters in $C_l^{\mathcal{P}}$. This can be formulated as follows:

$$a_{i,l}^{(d)} = \langle a_i^{(d-1)}, \mathcal{C}_{i,l}^{(d-1)} \rangle, \mathcal{C}_{i,l}^{(d-1)} = \left\{ C_{l,x}^{\mathcal{P}} \;\mid\; p_a^{(d-1)} \in C_{l,x}^{\mathcal{P}}, w_{i,a} \neq 0, \forall p_a, \forall C_{l,x}^{\mathcal{P}} \right\}, \tag{4}$$

where $\mathcal{C}_{i,l}^{(d-1)}$ indicates a list of clusters that are adjacent to $a_i^{(d-1)}$ at a detail level of l, and $C_{l,x}^{\mathcal{P}}$ refers to the xth cluster of subgraphs rooted in papers at a detail level of l.

Algorithm 5. WL relabeling process with levels of detail.

```
procedure WLreLABELLING$\mathcal{N}, \mathcal{C}$
    for $d : 1 \to D$ do
        for $l : 1 \to L$ do
            for $a_i \in \mathbb{A}$ do
                $\mathcal{C}_{i,l}^{(d-1)} \longleftarrow \{C_{a,l}^{\mathcal{P}} \mid p_{a,l}^{(d-1)} \in C_{i,l}^{\mathcal{P}}, w_{i,a} \neq 0, \forall p_a, \forall C_{a,l}^{\mathcal{P}}\}$
                $a_{i,l}^{(d)} \longleftarrow \langle a_i^{(d-1)}, \mathcal{C}_{i,l}^{(d-1)} \rangle$
                Put $\left\{ HASH \left( a_{i,l}^{(d)} \right) : a_{i,l}^{(d)} \right\}$ into $\mathcal{C}$
            end for
            for $p_a \in \mathbb{P}$ do
                $\mathcal{C}_{a,l}^{(d-1)} \longleftarrow \{C_{i,l}^{\mathcal{A}} \mid a_{i,l}^{(d-1)} \in C_{i,l}^{\mathcal{A}}, w_{i,a} \neq 0, \forall a_{i,l}, \forall C_{i,l}^{\mathcal{A}}\} \cup$
                $\{C_{n,l}^{\mathcal{J}} \mid j_{n,l}^{(d-1)} \in C_{n,l}^{\mathcal{J}}, p_{a,n} \neq 0, \forall j_{n,l}, \forall C_{n,l}^{\mathcal{J}}\}$
                $p_{a,l}^{(d)} \longleftarrow \langle p_a^{(d-1)}, \mathcal{C}_{a,l}^{(d-1)} \rangle$
                Put $\left\{ HASH \left( p_{a,l}^{(d)} \right) : p_{a,l}^{(d)} \right\}$ into $\mathcal{C}$
            end for
            for $j_n \in \mathbb{J}$ do
                $\mathcal{C}_{n,l}^{(d-1)} \longleftarrow \{C_{a,l}^{\mathcal{P}} \mid p_{a,l}^{(d-1)} \in C_{a,l}^{\mathcal{P}}, p_{a,n} \neq 0, \forall p_a, \forall C_{a,l}^{\mathcal{P}}\}$
                $j_{n,l}^{(d)} \longleftarrow \langle j_{n,l}^{(d-1)}, \mathcal{C}_{n,l}^{(d-1)} \rangle$
                Put $\left\{ HASH \left( j_{n,l}^{(d)} \right) : j_{n,l}^{(d)} \right\}$ into $\mathcal{C}$
            end for
        end for
    end for
end procedure
```

The preliminary representations of entities are generated using Subgraph2Vec (Narayanan et al., 2016), which is based on the Skip-Gram and negative sampling in Word2Vec (Mikolov et al., 2013). The objective function and neighborhoods (context in Word2Vec) for learning representations can be formulated as follows:

$$\mathcal{L}\left(a_i^{(d)}\right) \simeq \sum_{\forall s_a \in \mathcal{N}_i^{(d)}} \log\sigma\left(\Phi(s_a)^{\top}\Phi\left(a_i^{(d)}\right)\right) + \sum_{j=1}^{k} \mathbb{E}_{s_b \sim P_n(\mathcal{S})}\left[\log\sigma\left(-\Phi(s_b)^{\top}\Phi\left(a_i^{(d)}\right)\right)\right], \tag{5}$$

$$\mathcal{N}_i^{(d)} = \left\{ p_a^{(d+\triangle d)} \;\mid\; w_{i,a} \neq 0, |\triangle d| \leq \mathcal{W}_D, \forall p_a \right\}, \tag{6}$$

where $P_n(\mathcal{S}) \propto U(\mathcal{S})^{\alpha}$ denotes a noise distribution of subgraphs ($\alpha \in [0, 1]$), $U(\mathcal{S})$ refers to a unigram distribution of subgraphs, k indicates the number of negative samples, and $\mathcal{N}_i^{(d)}$ refers to neighborhoods of $a_i^{(d)}$. In addition, s_a indicates the ath subgraph in our dictionary, $\Phi(\cdot)$ denotes the projection function, and $\sigma(\cdot)$ refers to the sigmoid function. As shown in Fig. 1, bibliographic entities have excessively skewed distributions. Thus, we set α to nearly 0; in addition, the hyper-parameter search procedures are as presented in Section 4, including $D, L, k, and\alpha$. The same method is also used for papers and venues. Since Subgraph2Vec is a well-known model, we do not present a detailed explanation herein to avoid redundancy.

3.3. Learning representations of research patterns

We propose a model for the multi-resolution representational learning of research patterns by extending the existing models of Lee and Jung (2019), Lee and Jung, (2020a,b), Lee et al. (2020), and Narayanan et al. (2016). Their models are modifications of the Skip-Gram method, which mainly extends the ranges of neighborhoods, as shown in Eq. (6). Narayanan et al. proposed a radial Skip-Gram method in Subgraph2Vec to make the size of neighborhoods flexible and consider the degrees of the subgraphs. Lee et al. added dynamic changes in the subgraphs to the radial Skip-Gram method. With a similar approach, we modify the radial Skip-Gram method to consider the detail levels, as indicated in Eq. (3).

The remaining parts of the proposed model are similar to the conventional Skip-Gram method and negative sampling. For the research pattern, we predict patterns of adjacent entities (the first condition in Eq. (3)), for e.g., the types of papers the scholar writes. In addition, we also predict the meanings of the research pattern for other degrees and detail levels (the second and third conditions in Eq. (3)). As the degree increases, the research patterns show more subdivided information. For example, research patterns of a_i at a degree of 0 denote the node types, a degree of 1 merely shows lists of papers, and a degree of 1 covers co-authorship and venues. In addition, from a detail level of $L+1$ to 0, we gradually simplify the publication list of a_i. Using the prediction results, we update the vector representations of the research patterns.

Given a research paper, our objective function maximizes the co-occurrence probability of the neighborhoods (Eq. (3)), and minimizes the probability of research patterns that are not in the neighborhoods. At the detail level $L+1$, the objective function is mostly similar to the radial Skip-Gram in Eq. (5). This is formulated as follows:

$$\mathcal{L}\left(a_{i,l}^{(d)}\right) = \sum_{\forall \mathcal{S}_a \in \mathcal{N}_{i,l}^{(d)}} \log P\left(\mathcal{S}_a | \Phi\left(a_{i,l}^{(d)}\right)\right) - \sum_{\forall \mathcal{S}_b \notin \mathcal{N}_{i,l}^{(d)}} \log P\left(\mathcal{S}_b | \Phi\left(a_{i,l}^{(d)}\right)\right) \tag{7}$$

$$\simeq \sum_{\forall \mathcal{S}_a \in \mathcal{N}_{i,l}^{(d)}} \log \sigma\left(\Phi(\mathcal{S}_a)^{\top} \Phi\left(a_{i,l}^{(d)}\right)\right) + \sum_{j=1}^{k} \mathbb{E}_{\mathcal{S}_b \sim P_n(\mathcal{S})}\left[\log \sigma\left(-\Phi(\mathcal{S}_b)^{\top} \Phi\left(a_{i,l}^{(d)}\right)\right)\right].$$

This objective function brings $\Phi(\mathcal{S}_a)$ and $\Phi(\mathcal{S}_b)$ close to each other when $\mathcal{S}_a$ and $\mathcal{S}_b$ are in the same neighborhood. Otherwise, it keeps them distant. As with the preliminary representations, we set $P_n(\mathcal{S})$ to a nearly uniform distribution ($\alpha \simeq 0$).

The proposed model generates research patterns and their vector representation for each degree and detail level. Thus, each bibliographic entity has vectors with a size of $(L+2) \times (D+1)$. According to existing studies (Mikolov et al., 2013; Narayanan et al., 2016), concatenating all vectors is the most effective way to represent the entities. However, in this study, the number of subgraphs makes the representation of the entities highly dimensional. In addition, averaging all representations causes excessive information loss. To solve this problem, we suggest three methods for combining the research pattern vectors into entity vectors: (i) Graph2Vec (Narayanan et al., 2017), (ii) the average for each detail level (DetailMean), and (iii) the average for each degree (DegreeMean).

Doc2Vec (Le & Mikolov, 2014) is an effective model for aggregating word vectors into document vectors. Graph2Vec (Narayanan et al., 2017) is a model used to apply the distributed bag-of-words version of the paragraph vector (PV-DBOW) method in Doc2Vec on a graph. Using Graph2Vec, we can observe the entity vectors from the occurrence of research patterns in each entity.

We can employ strong points of both the average and concatenation. DetailMean averages the research pattern vectors at each detail level and concatenates the averaged vectors. In contrast, DegreeMean applies the average for each degree. DetailMean and DegreeMean respectively emphasize the degrees and detail levels more than the other. These methods can be formulated as follows:

$$\Phi(a_i) = oplus_{0 \le d \le D} \underset{0 \le l \le L+1}{mean} \Phi\left(a_{i,l}^{(d)}\right), \Phi(a_i) = oplus_{0 \le l \le L+1} \underset{0 \le d \le D}{mean} \Phi\left(a_{i,l}^{(d)}\right), \tag{8}$$

where $\oplus$ indicates the concatenation operator. In Section 4.3, we present an empirical experiment for deciding the adequate combination method for bibliographic entities.

Table 1
Statistics of the bibliographic network. # Scholars, # Papers, and # Venues are the numbers of scholars, papers, and venues in our bibliographic network, respectively. # Nodes and # Edges indicate the total number of nodes and edges in our network, respectively. # Collab. and # Citations denote the numbers of edges in the co-authorship and citation networks, respectively.

Periods	# Scholars	# Papers	# Venues	# Nodes	# Edges	# Collab.	# Citations
2011–2015	260,726	196,595	141	457,462	661,700	768,092	304,658
2011–2016	306,665	248,661	141	555,467	789,373	1,015,532	457,087
2011–2017	353,880	304,245	141	658,266	923,443	1,195,619	646,976
2011–2018	396,857	351,914	141	748,912	1,041,875	1,364,067	827,209

4. Evaluation

This section aims at evaluating the proposed methods and validating the research questions (in Section 1). We examined the effectiveness of the proposed representations in predicting the 23 performance indicators during four time periods. If the research pattern vectors are useful for predicting diverse aspects of the performance during all periods, (i) the bibliographic entities have their own research styles, which are more consistent than the performance at each moment, and (ii) the proposed representations can reflect the styles (RQ 1). In addition, we verified the robustness of the prediction accuracy on the performance groups of the bibliographic entities. This experiment presents the possibility of the proposed model in improving the imbalanced learning opportunities between the high- and low-performance entities (RQ 2).

4.1. Data collection and implementation

We composed the bibliographic network by refining the ArnetMiner dataset[4] (Sinha et al., 2015), which has been widely used as a benchmark dataset to evaluate network embedding models. This dataset consists of bibliographic data collected from DBLP, the ACM Digital Library,[5] Microsoft Academic Graph,[6] and other sources. Sinha et al. (2015) have been distributing collected data online after a preprocessing process, including a disambiguation of the bibliographic entities' identities. However, for scholars and venues, not all of their papers were collected in a consistent fashion. Thus, from the latest version of the ArnetMiner dataset (DBLP-Citation-network V11), we obtained papers published from 2011 to 2018 and removed the venues and scholars with incomplete bibliography data. The original version consists of 4,107,340 papers and 36,624,464 citation relationships. In addition, our refined version includes 351,914 papers published by 396,857 scholars in 141 venues. Since not all the pre-existing embedding models are for heterogeneous networks, we also composed a co-authorship and citation network of papers and venues with the same dataset. Table 1 presents detailed statistics of the collected bibliographic network. Furthermore, the proposed embedding model was implemented based on open-source implementations of Subgraph2Vec[7] ,[8] and Collaboration2Vec.[9] The implementation of the proposed model has also been made available through an open-source repository.[10]

Moreover, the proposed methods require various hyper-parameters. In particular, the maximum detail level L, the maximum degree D, and the window size for detail level $\mathcal{W}_L$ and degree $\mathcal{W}_D$ affect both the research pattern discovery and the representation learning of the patterns. Thus, we conducted a grid search for these parameters: L (5 to 11 with a step size of +2), D (5 to 11 with a step size of +2), $\mathcal{W}_L$ (1 to 3 with a step size of +1), and $\mathcal{W}_D$ (1 to 3 with a step size of +1). Each case was evaluated based on the mean square error (MSE, also called L2 norm) for predicting the h-index of the scholars; the experimental procedures are described in the next section. The proposed model performed the best at $L = 7$, $D = 7$, $\mathcal{W}_L = 2$, and $\mathcal{W}_D = 2$. In addition, we must determine the hyper-parameters that come from the negative sampling and the Skip-Gram method. However, this study focuses on enabling the multi-resolution bibliographic network embedding rather than enhancing its accuracy. Therefore, we determined the parameters in a heuristic manner; the number of epochs (ε) was set to 200, the learning rate (η) was 0.0025, the number of dimensions (δ) was 64, the number of negative samples (k) was 100, and the weighting factor for noise distribution (α) was 0.0. For example, because the bibliographic entities are highly imbalanced (Fig. 1), α should be small for representing the outstanding entities.

4.2. Experimental procedures

We compared the proposed and existing embedding models by predicting 23 quantitative indicators for bibliographic entities (Table 2) by using their vector representations and neural networks. The performance of a scholar was mainly measured based on the number of papers (# papers) and the number of citations (# citations). To consider both of them,

[4] https://www.aminer.cn/citation.
[5] https://dl.acm.org/.
[6] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[7] https://github.com/MLDroid/subgraph2vec_tf.
[8] https://github.com/MLDroid/subgraph2vec_gensim.
[9] https://github.com/higd963/Collaboration2Vec.
[10] https://github.com/higd963/Multi-resolution-Network-Embedding.

Table 2
List of quantitative indicators. The circles depict the usage of the indicators for each entity type. The acronyms used are as follows; # of papers indicates the number of papers written by the scholars, # of citations indicates the number of citations for their papers, $\bar{\#}$ of citations is the average number of citations for each paper, J/C is the ratio of the number of journal articles to the number of conference papers, PR indicates the PageRank centrality, BC is the betweenness centrality, CC is the closeness centrality, IF is the impact factor, ES is the eigenfactor score, and AS indicates the article score.

Quantitative	Entity types		
Indicators	Scholars	Papers	Venues
h-index (Hirsch, 2005)	○		○
# Papers	○		○
# Citations	○	○	○
$\bar{\#}$ Citations	○		○
J/C	○		
IF (Garfield, 2006)		○	○
ES (Bergstrom et al., 2008)			○
AS (Bergstrom et al., 2008)			○
PR (Page, Brin, Motwani, & Winograd, 1999)	○	○	○
BC (Freeman, 1977)	○	○	○
CC (Sabidussi, 1966)	○	○	○

we measured the *h*-index (Hirsch, 2005) and the average number of citations ($\bar{\#}$ citations). In addition, the ratio of journal articles for conference papers (J/C) will show the preferences of the scholars. For the papers, we used # of citations and the journal impact factor (IF) (Garfield, 2006) for their venues. Finally, the venue quality was measured based on # of papers and # of citations. To consider them together, we used the *h*-index, $\bar{\#}$ citations, IF, eigenfactor score (ES), and article score (AS) (Bergstrom et al., 2008). Moreover, a few existing studies (Mariani, Medo, & Zhang, 2016; Waheed, Imran, Raza, Malik, & Khattak, 2019; Ye, Li, & Law, 2011) evaluated scholars and papers using their node centrality on the co-authorship networks and citation networks. We also composed the citation network for the venues, and the PageRank (PR), betweenness (BC), and closeness centrality (CC) were commonly applied to the three entity types.

In 2020, we cannot predict the number of papers a scholar will write in 2025, the co-authorship, or the venue. However, if the proposed representations reflect the research styles of the bibliographic entities, we will be able to predict future academic activities of the entities. In addition, the performance of the bibliographic entities have various aspects. Some quantitative indicators may reflect the aspects immediately, whereas others may have time delays. For example, citation-based indicators (e.g., # citations and *h*-index) will have a delay of 2 or 3 years as the papers are supposed to be read by other scholars, and the readers need time to write and publish their papers. Therefore, we generated the vector representations considering the bibliographical data from 2011 to 2015. We then examined the prediction of the future performance of the entities by using the representations. The indicators were measured for the recent five years. Because we collected data from 2011 to 2018, we have four sets of indicators: 2011 to 2015, 2011 to 2016, 2011 to 2017, and 2011 to 2018.

To validate the research questions, we compared the accuracy of the proposed model with the existing network embedding model DeepWalk (Perozzi et al., 2014), LINE (Tang et al., 2015),[11] Metapath2Vec (Dong et al., 2017),[12] HIN2Vec (yang Fu et al., 2017),[13] and Subgraph2Vec (Narayanan et al., 2016),[14] and Collaboration2Vec (Jeon et al., 2019). Among the existing models, DeepWalk, LINE, and Collaboration2Vec do not consider the heterogeneity of the nodes or relations. Among them, Collaboration2Vec is designed only for a co-authorship network. Thus, in this case, we applied a co-authorship network for a fair comparison. The hyper-parameters of the existing models followed the open-source implementations of their authors. In addition, we empirically tuned those parameters that are yet to be determined in the official implementations.

For the prediction, we trained the fully connected neural networks having five hidden layers. The inputs of the neural networks were vector representations generated by the proposed and existing embedding models. The outputs of the neural networks were quantitative indicators for the bibliographic entities. All layers used ReLu (Rectified Linear Unit) for their activation function. The numbers of neurons on each layer were as follows: input layer, 576; hidden layer 1, 512; hidden layer 2, 256; hidden layer 3, 128; hidden layer 4, 64; hidden layer 5, 32; and output layer, 1. The loss function and evaluation metric commonly applied the MSE.

Furthermore, we expected that the proposed embedding model can obtain more distinctive representations of outstanding bibliographic entities than the existing models. Thus, we compared the prediction performance according to the performance and quality of the entities by composing four groups for each type of entity: top 1%, top 5%, top 10%, and the rest. Since the rankings of the entities were different for each indicator, we composed the groups considering each indicator. In addition, owing to the small number of venues (only 141), we could not group them. In summary, we examined the 7 embedding models for the 23 indicators on the 4 time periods with the 4 performance groups. We built and trained independent neural networks for each case.

[11] https://github.com/shenweichen/GraphEmbedding.
[12] https://github.com/apple2373/metapath2vec
[13] https://github.com/csiesheep/hin2vec.
[14] This method is similar to the preliminary vector representations in Eq. (5), as described in Section 3.2.

Table 3
Accuracy for predicting the top 10% of the scholars and papers according to the concatenation methods applied. Since the indicators of the scholars and papers have a highly skewed distribution, we assessed the concatenation methods using only the top 10% of the papers.

Years	2015	2016	2017	2018	Average
Graph2Vec	0.22	0.32	0.42	0.49	0.37
DetailMean	0.12	0.28	0.31	0.43	0.26
DegreeMean	**0.11**	**0.18**	**0.23**	**0.29**	**0.19**

Table 4
Accuracy of the proposed and existing embedding models for predicting the performance of the scholars. Cells depict the average normalized MSE for each case. The values in the round brackets denote the standard deviation of the normalized MSE.

		Proposed	Subgraph2Vec	Collaboration2Vec	DeepWalk	LINE	Metapath2Vec	HIN2Vec
2015	100%	**0.11**** (**0.13****)	0.26 (0.30)	0.35 (0.23)	0.12 (0.17)	0.35 (0.20)	0.26 (0.19)	**0.05*** (**0.06***)
	10%	**0.16*** (**0.18****)	**0.20**** (0.30)	0.26 (0.20)	0.28 (0.32)	0.48 (**0.16***)	0.46 (0.20)	0.24 (0.28)
	5%	**0.14**** (0.21)	0.30 (0.36)	0.22 (0.27)	**0.12*** (**0.15***)	0.36 (**0.20****)	0.32 (0.27)	0.19 (0.29)
	1%	**0.06*** (**0.10***)	**0.11**** (**0.11****)	0.22 (0.21)	0.18 (0.12)	0.33 (0.20)	0.27 (0.28)	0.12 (0.14)
2016	100%	**0.10**** (**0.07***)	0.23 (0.25)	0.50 (0.25)	0.20 (0.25)	0.41 (0.22)	0.36 (0.19)	**0.10*** (**0.08****)
	10%	**0.15*** (0.19)	0.35 (0.32)	0.31 (**0.12***)	**0.27**** (0.21)	0.59 (0.19)	0.71 (**0.16****)	0.29 (0.22)
	5%	**0.15*** (**0.14***)	**0.18**** (0.23)	0.23 (0.29)	0.28 (0.32)	0.49 (**0.21****)	0.40 (0.21)	0.27 (0.28)
	1%	**0.13*** (**0.11****)	**0.15**** (0.09)	0.31 (0.25)	0.33 (0.23)	0.34 (0.20)	0.20 (**0.07***)	0.18 (0.19)
2017	100%	**0.20*** (**0.09***)	0.33 (0.27)	0.73 (0.27)	0.32 (0.29)	0.53 (0.29)	0.43 (0.29)	**0.26**** (**0.27****)
	10%	**0.20*** (**0.12***)	0.26 (**0.15****)	0.51 (0.25)	**0.25**** (0.12)	0.66 (0.25)	0.82 (0.16)	0.35 (0.23)
	5%	**0.24**** (0.20)	**0.22*** (**0.13***)	0.29 (0.27)	0.33 (0.23)	0.75 (**0.18****)	0.72 (0.21)	0.24 (0.19)
	1%	**0.17*** (**0.06***)	0.21 (0.12)	0.37 (0.18)	0.38 (0.25)	0.53 (0.22)	0.35 (0.28)	**0.17**** (**0.12****)
2018	100%	**0.19*** (**0.07***)	0.33 (0.26)	0.60 (0.14)	**0.22**** (**0.11****)	0.61 (0.31)	0.67 (0.38)	0.40 (0.29)
	10%	**0.13*** (**0.16***)	0.24 (0.30)	0.28 (0.27)	0.23 (0.26)	0.31 (**0.20****)	0.41 (0.31)	**0.18**** (0.25)
	5%	**0.23*** (**0.13***)	0.44 (0.19)	**0.33**** (**0.16****)	0.34 (0.09)	0.81 (0.22)	0.60 (0.31)	0.42 (0.24)
	1%	**0.30*** (**0.10***)	0.42 (0.26)	0.56 (0.28)	0.41 (**0.20****)	0.61 (0.40)	0.49 (0.31)	**0.39**** (0.29)
Avg.	100%	**0.15*** (**0.10***)	0.29 (0.27)	0.55 (0.27)	0.21 (**0.23****)	0.47 (0.28)	0.43 (0.31)	**0.20**** (0.25)
	10%	**0.16*** (**0.17***)	0.26 (0.28)	0.34 (**0.24****)	**0.25**** (0.24)	0.52 (0.24)	0.59 (0.28)	0.26 (0.25)
	5%	**0.19*** (**0.18***)	0.28 (0.26)	**0.27**** (0.25)	0.27 (0.23)	0.60 (**0.27****)	0.51 (0.30)	0.28 (0.27)
	1%	**0.17*** (**0.13***)	**0.22**** (**0.20****)	0.37 (0.26)	0.33 (0.23)	0.45 (0.30)	0.33 (0.28)	0.22 (0.22)

4.3. Experimental results and discussion

The proposed model embeds each research pattern, and the bibliographic entities have $(L+2)\times(D+1)$ research patterns. To compose the representations of each entity, we presented three methods for combining the research pattern vectors, namely, Graph2Vec, DetailMean, and DegreeMean. Prior to evaluating the proposed model, we had to determine the method of combination. Because this study aims at achieving outstanding entities, Table 3 presents the accuracy of the combined vectors for predicting the quantitative indicators of the top 10% of the bibliographic entities. We measured the accuracy by using the MSE between the predicted and actual indicator values. Because the range of indicator values is diverse, we normalized the MSE for each indicator. Table 3 shows the average value of the normalized MSE.

DegreeMean achieved the highest accuracy consistently throughout all the periods. This result indicates that the detail levels of the research patterns are more effective for representing outstanding entities than the degrees. There was a significant gap between the two concatenation methods and Graph2Vec. Graph2Vec might excessively simplify information in the research pattern vectors. Using DegreeMean, we compared the proposed model with the existing embedding models according to the performance groups and periods.

Tables 4 , 5 , and 6 present the accuracy of the proposed and existing models for the scholars, papers, and venues, respectively. We assessed the accuracy for predicting the 23 quantitative indicators. We normalized the MSE for each indicator and calculated the arithmetic mean and standard deviation of the normalized errors for each case. The raw MSE data before the normalization is accessible through the open source repository.[15]

As shown in Table 4, the proposed model exhibited the highest accuracy for most of the cases in terms of both the average and variance of the MSE. On an average, the proposed model achieved the best accuracy for all performance groups. This result indicates that the level-wise simplification makes the embedding model robust to skewed distributions of bibliographic entities. However, considering the years 2015–2017, the proposed model showed the second-highest accuracy in a few cases (100% and 5%). A total of 1% of the cases exhibited the effectiveness of the proposed model for outstanding scholars. As shown in Fig. 1, the performances of bibliographic entities are highly skewed. Thus, if we assume all the indicators to be 0, we will obtain a high accuracy for 100% of the cases. Although the top 10%, 5%, and 1% of entities are roughly skewed, the excessive diversity begins near the top 0.1% of entities. We can assume that the strategy for 100% of the cases still worked on 10% of the cases; moreover, 1% of the cases showed the ability to handle the diversity of the outstanding entities; however,

[15] https://github.com/higd963/Multi-resolution-Network-Embedding/tree/master/03_results.

Table 5
The accuracy of the proposed and existing embedding models for predicting the performance of the papers.

		Proposed	Subgraph2Vec	DeepWalk	LINE	Metapath2Vec	HIN2Vec
2015	100%	**0.17*** (**0.18***)	0.24 (0.22)	**0.19**** (**0.19****)	0.34 (0.20)	0.32 (0.32)	0.20 (0.19)
	10%	**0.17*** (0.21)	**0.18**** (0.22)	0.24 (0.32)	0.21 (**0.20****)	0.22 (**0.14***)	0.35 (0.42)
	5%	**0.15*** (0.18)	0.38 (0.43)	0.32 (0.28)	0.16 (**0.15****)	0.29 (0.26)	**0.16**** (**0.14***)
	1%	**0.18**** (0.23)	**0.18*** (0.22)	0.20 (0.23)	0.19 (**0.19***)	0.19 (**0.21****)	0.21 (0.30)
2016	100%	0.22 (0.21)	0.20 (0.21)	**0.18**** (**0.19****)	0.27 (**0.15***)	0.22 (0.20)	**0.14*** (0.21)
	10%	**0.25*** (0.19)	0.31 (0.20)	**0.25**** (0.17)	0.30 (**0.12***)	0.50 (**0.15****)	0.28 (0.27)
	5%	**0.18**** (**0.14****)	0.20 (0.19)	0.33 (0.30)	0.31 (0.18)	0.45 (0.36)	**0.14*** (**0.10***)
	1%	**0.17*** (**0.12***)	0.30 (0.28)	0.22 (0.19)	0.33 (0.30)	0.23 (**0.13****)	**0.20**** (0.20)
2017	100%	0.27 (0.24)	0.30 (**0.17****)	**0.16*** (**0.10***)	0.61 (0.33)	**0.24**** (0.20)	0.48 (0.43)
	10%	**0.27*** (**0.17****)	0.41 (0.24)	**0.35**** (0.20)	0.54 (**0.12***)	0.73 (0.24)	0.42 (0.23)
	5%	0.32 (0.19)	**0.13*** (**0.13****)	**0.22**** (**0.12***)	0.54 (0.20)	0.47 (0.39)	0.24 (0.17)
	1%	**0.28*** (**0.20****)	0.41 (0.26)	0.49 (0.25)	0.59 (0.35)	**0.34**** (**0.15***)	0.40 (0.33)
2018	100%	**0.27*** (**0.17***)	0.51 (0.27)	0.41 (**0.18****)	0.56 (0.32)	0.53 (0.23)	**0.35**** (0.21)
	10%	**0.25*** (**0.20***)	0.44 (0.33)	**0.32**** (0.26)	0.51 (0.33)	0.65 (**0.20****)	0.43 (0.30)
	5%	**0.36*** (0.32)	0.59 (0.34)	0.55 (0.39)	0.54 (**0.31****)	**0.43**** (**0.27***)	0.44 (0.36)
	1%	**0.36**** (0.22)	0.45 (0.32)	**0.32*** (0.38)	0.68 (0.31)	0.82 (**0.12***)	0.42 (**0.20****)
Avg.	100%	**0.23*** (**0.20****)	0.31 (0.25)	**0.23**** (**0.20***)	0.44 (0.30)	0.33 (0.27)	0.29 (0.31)
	10%	**0.24*** (**0.20****)	0.34 (0.27)	**0.29**** (**0.25***)	0.39 (0.25)	0.47 (0.28)	0.37 (0.33)
	5%	**0.25**** (**0.23***)	0.32 (0.34)	0.36 (0.31)	0.39 (0.27)	0.41 (0.33)	**0.24*** (**0.25****)
	1%	**0.25*** (**0.21****)	0.33 (0.29)	0.31 (0.30)	0.45 (0.35)	0.39 (0.30)	**0.31**** (**0.28***)

Table 6
Accuracy of the proposed and existing embedding models for predicting venue performance.

	Proposed	Subgraph2Vec	DeepWalk	LINE	Metapath2Vec	HIN2Vec
2015	**0.08*** (**0.12***)	0.17 (0.24)	**0.13**** (**0.19****)	0.21 (0.20)	0.32 (0.34)	0.20 (0.25)
2016	**0.08*** (**0.14***)	**0.19**** (**0.24****)	0.19 (0.25)	0.45 (0.37)	0.31 (0.34)	0.32 (0.36)
2017	**0.05**** (**0.08****)	0.32 (0.37)	**0.03*** (**0.04***)	0.28 (0.27)	0.42 (0.35)	0.12 (0.18)
2018	**0.10*** (**0.13***)	0.24 (0.26)	**0.10**** (**0.14****)	0.42 (0.34)	0.28 (0.26)	0.23 (0.27)
Avg.	**0.08*** (**0.12***)	0.23 (0.28)	**0.11**** (**0.16****)	0.34 (0.29)	0.32 (0.33)	0.22 (0.26)

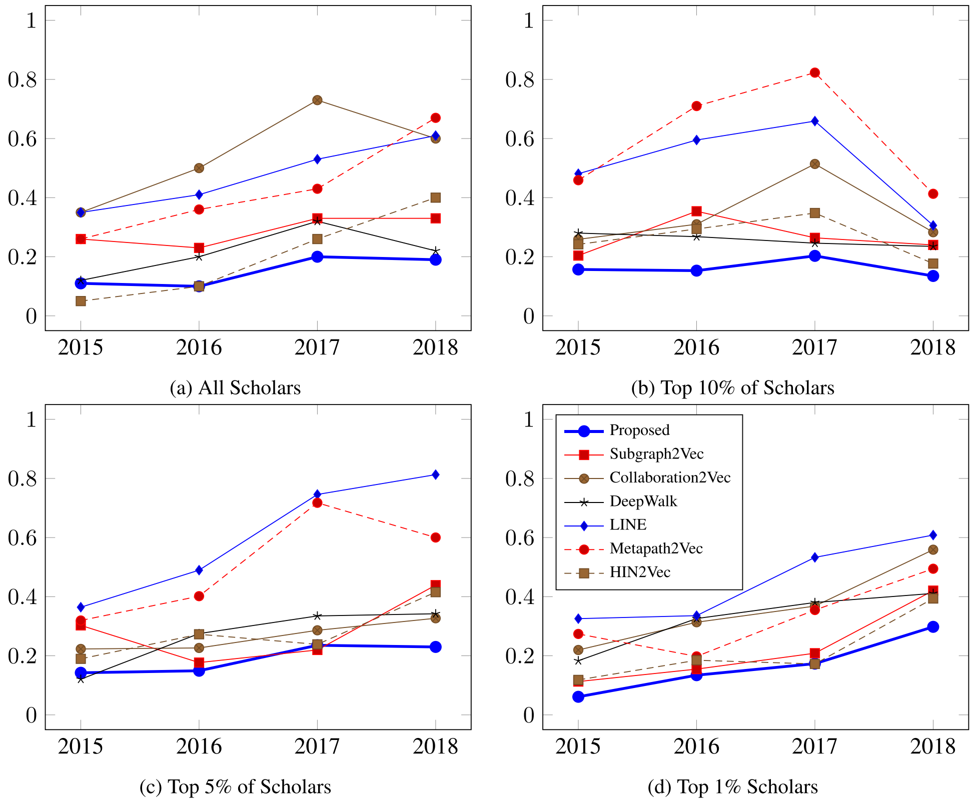
the two points had conflicts in 5% of the cases. We could not find the exact reasons for this, and we will attempt to examine them in a future study.

Among the existing models, Subgraph2Vec, DeepWalk, and HIN2Vec exhibited a high accuracy (nearby 0.20). In addition, Collaboration2Vec achieved a high accuracy, excluding 100% of the cases. Although Metapath2Vec showed a similar accuracy as DeepWalk for 1% of the cases, LINE and Metapath2Vec had a commonly low accuracy for the other cases. The accuracy of LINE underpins the fact that substructure-based embedding models are more suitable for a bibliographic network than proximity-based models. For Metapath2Vec, the symmetric metapaths employed by this model might not be appropriate for representing the research styles of the scholars. The accuracy of Collaboration2Vec and DeepWalk were relatively lower than those of the proposed model, Subgraph2Vec, and HIN2Vec. This result indicates the necessity of the heterogeneity in bibliographic network embedding.

The proposed model, Subgraph2Vec, and Collaboration2Vec are all based on subgraphs discovered through the WL relabeling process. In addition, DeepWalk, Metapath2Vec, and HIN2Vec use random walkers. However, excluding the best and worst cases (the proposed model and Metapath2Vec), the remaining four methods did not show differences in the types of substructures. All of the embedding models are based on the Skip-Gram method, unlike the HIN2Vec and LINE. The learning methods of HIN2Vec are more sophisticated than those of the Skip-Gram method. However, HIN2Vec could not distinctly outperform the other models, and was outperformed by the proposed model. Methods for composing the substructures might have more contributions to the accuracy than the learning methods. Finally, Collaboration2Vec achieved a reasonable level of accuracy, although this model considers only subgraphs in co-authorship networks. In addition, Collaboration2Vec showed a particularly high performance for outstanding scholars, which is incredibly diverse. As we assumed, the information on the scholars, the papers they published, the venues in which the papers were published, and their collaborators, represent the research styles of the scholars.

In Fig. 6, we can observe changes in the accuracy according to the periods and performance groups. Most of the models achieved a better performance in short-term prediction than in long-term prediction. However, the proposed model performed relatively slower in terms of reduction than the other methods. The proposed model achieved a similar accuracy as the existing models for the short-term prediction. On the other hand, in the long-term cases, the proposed model distinctly outperformed the others, even for 100% and 5% of the cases. The level-wise simplification and multi-resolution embedding were effective in discovering the research styles of the scholars, which are not merely temporary.

Nevertheless, not all of the models achieved a gradual performance reduction in accordance with the years. Many of the models showed a higher accuracy for the papers published in 2018 than for 2017, particularly for the top 10% scholars. Even LINE achieved its highest accuracy within the top 10% in 2018. This likely occurred for the following two reasons. First, most of the indicators are related to citations, and there will be time delays between publications and their citations. Second,



(a) All Scholars

(b) Top 10% of Scholars

(c) Top 5% of Scholars

(d) Top 1% Scholars

Fig. 6. Normalized MSE of the proposed and existing embedding models for predicting the performance of the scholars according to the time periods: (a)–(d) present changes in the accuracy of the embedding models according to performance groups. The *X*- and *Y*-axes of the plots indicate the years and MSE, respectively.

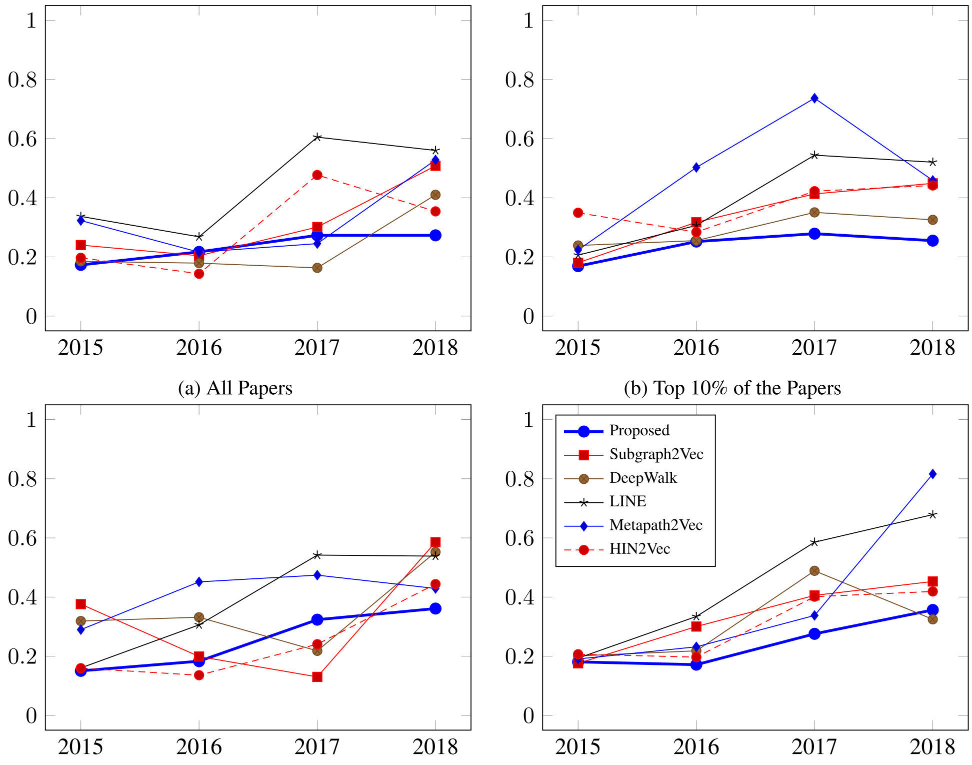
talented researchers might become more distinguished after 3 years. The top 5% and 1% of the scholars might already be senior researchers, and the rookies will most likely be in the top 10% group.

Table 5 shows the experimental results for the papers. As in the previous case, the proposed model exhibited the best performance in most of the cases. Nevertheless, its weakness for the 100% and 5% cases was more severe for the papers than for the scholars. Even for the top 1% of the papers, the proposed model achieved the second highest accuracy for the papers published in 2015 and 2018. In terms of the variance, the proposed model showed the second-lowest variance overall. This result underpins the effectiveness of the level-wise simplification method.

This result might come from two points. First, papers are less connected with the other nodes compared to the scholars and venues. Scholars and venues have links with multiple papers. In addition, the papers provide abundant explanation. However, a paper is connected to a limited number of authors and a single venue. Although the WL relabeling process extends the coverage of the subgraphs according to the degree, a small number of connections might restrict the diversity of the subgraphs. Comparing the performance of Subgraph2Vec for scholars with that of papers underpins this assumption. Second, the first-order proximity of the papers, which is different from the scholars and venues, does not change according to time. The research styles that the proposed model aimed at might not be significant for the papers. For scholars and venues, we have to predict further changes in their substructures; however, the papers do not require such a prediction.

The random-walk based models (DeepWalk, Metapath2Vec, and HIN2Vec) exhibited a better performance for papers than for scholars. The random walkers search for adjacent nodes having a probability distribution. For papers, a random walk might reflect broader structural information than a subgraph. The proximity-based approach (LINE) showed the worst performance for both scholars and papers. However, the accuracy was much better for papers than for scholars. Notably, LINE achieved the third-highest accuracy in 2015 for outstanding papers (1%, 5%, and 10%). The proximity was insufficient to describe the styles of the scholars and venues. However, since the papers are static, the second-order proximity might be sufficient to represent the similarity between papers with shared authors and venues.

As shown in Fig. 7, the proposed model showed a distinctly slower performance reduction than the existing models. In contrast, the other models demonstrated higher irregular tendencies for papers than for scholars. Network structures might be too detailed to discover correlations based on the research performance. We assume that the simplified structures can reveal the correlations by reducing the overabundant information. Therefore, the proposed model achieved the highest accuracy for the papers published in the year 2018, despite its low accuracy for those in 2016 and 2017. This stable accuracy



(a) All Papers

(b) Top 10% of the Papers

(c) Top 5% of the Papers

(d) Top 1% Papers

Fig. 7. Normalized MSE of the proposed and existing embedding models for predicting the performance of papers according to the year.

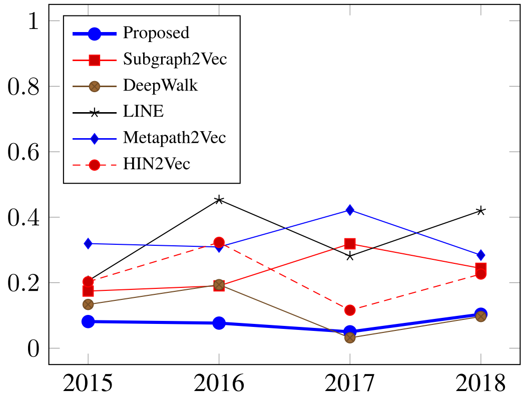


Fig. 8. Normalized MSE of the proposed and existing embedding models for predicting venue performance according to the years.

underpins the idea that multi-resolution embedding is currently better suitable for representing research styles than the performance levels.

Table 6 and Fig. 8 show experimental results for the venues. Because the number of venues is relatively smaller than the number of scholars and papers, we conducted the experiments exclusively for all the venues. Excluding 2017, the proposed model achieved the highest accuracy for all the periods. Differring from the other node types, the proposed model showed significant performance gaps in comparison with the other models for the short-term prediction (2015 and 2016). For the long-term, DeepWalk exhibited a similar performance as the proposed approach.

We can interpret this result mainly in three ways. First, publications affect the performance indicators of venues faster than it affects scholars. The scientific impact of the papers is not agile. For example, it could take a few months to a few years until a paper is cited by other authors. Therefore, the number of papers might have more influence on the indicators used for venues than on the indicators applied to scholars. Second, the subgraph-based models achieved a higher accuracy in the first 2 years, whereas the random walk-based models (excluding Metapath2Vec) showed a better performance in the last 2 years. The styles of the venues might not be as consistent as the styles of the scholars. For the advent of new research trends, conducting research in a novel way will be harder than collecting novel types of papers. Finally, subgraphs aim at a complete representation of a structure in terms of coverage. By contrast, random walks are broader and less detailed than the subgraphs. Because venues have a far greater number of adjacent nodes than scholars, the subgraphs rooted in venues can be excessively detailed.

Only DeepWalk was comparable with the proposed approach. This result is unexpected because DeepWalk does not consider the heterogeneity. The unexpectedness was common in all node types. Our bibliographic network uses a tripartite graph. This structural characteristic might offset the absence of heterogeneity. Subgraph2Vec and HIN2Vec make up the second-tier group. HIN2Vec has a more sophisticated learning method than Skip-Gram of DeepWalk. This model considers not only the adjacency of the nodes but also the directions and types of relations between nodes. However, in bibliographic networks, the relation is determined by its head and tail. Thus, the relation information can be excessive. As with the scholar cases, LINE and Metapath2Vec were among the worst groups, and their low accuracy might occur for the same reasons discussed for the scholars.

The proposed model clearly outperformed the existing models, particularly for scholars and venues. Based on a comparison with Subgraph2Vec, we validated the effectiveness of the detail level and multi-resolution representation learning in bibliographic network embedding. However, for the papers and venues, we also found strong points in the random walk approaches. In further studies, we will embrace the broad coverage of random walks and adjust the meticulousness of the subgraphs. We also plan to consider the significance of the nodes during WL relabeling.

5. Conclusion

This study aimed at representing the research patterns of bibliographic entities with a fixed-length vector. Outstanding bibliographic entities have diverse styles in comparison with their volume, and low-performance entities have far higher frequencies than those of outstanding entities. Thus, the existing models suffered from an overabundance of low-performance entities and a lack of samples for high-performance entities. We solved this problem by simplifying the research patterns (subgraphs rooted in the entities) according to multiple detail levels. Because the simplified patterns provide co-occurrences between infrequent and frequent patterns, we can balance the learning opportunities between high-performance and low-performance entities.

Our experiment validated that (i) the vector representations of the research patterns reflect the research styles of the bibliographic entities (RQ 1) and (ii) the proposed embedding model is robust to the skewed distribution of the entities (RQ 2). First, the proposed representations achieved a higher accuracy and lower variance for 23 quantitative indicators in comparison with the existing embedding models. This point underpins the idea that bibliographic entities have their own styles covering their various aspects. Second, the proposed model exhibited considerably lower performance reductions in long-term prediction than in existing models. Thus, although fluctuations may occur in a few performance aspects (e.g., the number of papers), the research styles of the entities are consistent. Finally, the proposed model outperformed the existing approaches for all performance groups. This result supports the idea that the proposed embedding model can resolve the imbalanced learning opportunities of the bibliographic entities. However, the proposed model has a few limitations that can be solved in future studies as follows:

- Adequate substructures for each entity type: Subgraphs were not the best choice for handling (i) the infrequent and static first-order proximity of the papers and (ii) the numerous adjacencies of the venues. For papers, we need substructures that have a broader coverage than subgraphs. In addition, venues require substructures that can consider the significance of the entities. We will attempt to apply a probabilistic random walk method based on the node centrality and employ multiple types of substructures for each entity type in further research.
- Temporal dynamics of bibliographic networks: The proposed model does not cover temporal changes in bibliographic networks. The high accuracy for long-term prediction implies that not only can the proposed model deal with the outstanding entities, but it can also distinguish promising researchers from other low-performance entities. Time-sequential changes will reinforce this strong point.

The proposed embedding model and research pattern vectors can be applied to various applications. We present directions for further research regarding its practical applications:

- Recommending and retrieving bibliographic entities: Representations of research styles will be effective for an academic team formation application (Yu et al., 2019). Working styles will be as important as the research capability in terms of choosing co-workers. Multi-resolution embedding also enables us to focus on bibliographic entities with particular styles or performances. This point will be useful in visualizing the research history of bibliographic networks (Shen et al., 2017) to filter overabundant information. Finally, we demonstrated that the research style can be used for a long-term performance prediction. Thus, the proposed representation can also be applied to measure the research capacity of a scholar and the prominence of the venues and research topics (Xiao et al., 2019).
- Building accurate bibliographic networks: There are academic databases available that provide relatively accurate bibliographic data, such as DBLP. However, their accuracy depends on restricting the collection range into selected venues or

particular research fields. Moreover, these databases suffer from several issues, such as the diversity of the reference styles, and changes in the names of the scholars or venues, among other issues. Accurate vector representations will improve the accuracy of the link prediction on bibliographic networks (Cai et al., 2019; Li et al., 2018; Zhou et al., 2017) and will help measure the trustworthiness of a citation or co-authorship relations. Finally, by measuring the similarity between research styles, we will obtain an additional feature for matching the identities of the scholars and venues.

Author contributions

O-Joun Lee: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Hyeon-Ju Jeon: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Jason J. Jung: Conceived and designed the analysis; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020R1A2B5B01002207). In addition, this research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Consilience Creative program (IITP-2019-2011-1-00783) supervised by the IITP (Institute for Information & communications Technology Planning & evaluation).

References

Abbasi, A., Altmann, J., & Hwang, J. (2009). Evaluating scholars based on their academic collaboration activities: Two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, *83*, 1–13, doi:10.1007/s11192-009-0139-2

Abramo, G., Cicero, T., & D'Angelo, C. A. (2013). Individual research performance: A proposal for comparing apples to oranges. *Journal of Informetrics*, *7*, 528–539. https://doi.org/10.1016/j.joi.2013.01.013

Anil, A., & Singh, S. R. (2020). Effect of class imbalance in heterogeneous network embedding: An empirical study. *Journal of Informetrics*, *14*, 101009. https://doi.org/10.1016/j.joi.2020.101009

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. *IEEE Access*, *7*, 9324–9339, doi:10.1109/access.2018.2890388.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The EigenfactorTM metrics. *Journal of Neuroscience*, *28*, 11433–11434, doi:10.1523/jneurosci.0003-08.2008.

Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, *9*, 135–144. https://doi.org/10.1016/j.joi.2014.12.001

Brand ao, M. A., Moro, M. M., Lopes, G. R., & Oliveira, J. P. (2013). Using link semantics to recommend collaborations in academic social networks. In L. Carr, A. H. F. Laender, B. F. Lóscio, I. King, M. Fontoura, D. Vrandecic, L. Aroyo, M. Palazzo, J. de Oliveira, F. Lima, & E. Wilde (Eds.), *Proceedings of the 22nd international conference on World Wide Web (WWW 2013)* (pp. 833–840). Rio de Janeiro, Brazil: ACM Press, doi:10.1145/2487788.2488058.

BV, E. (2018). *Research metrics guidebook*. Elsevier.

Cai, L., Tian, J., Liu, J., Bai, X., Lee, I., Kong, X., & Xia, F. (2018). Scholarly impact assessment: A survey of citation weighting solutions. *Scientometrics*, *118*, 453–478, doi:10.1007/s11192-018-2973-6.

Cai, X., Han, J., & Yang, L. (2018). Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation. In S. A. McIlraith, & K. Q. Weinberger (Eds.), *Proceedings of the 32nd AAAI conference on artificial intelligence (AAAI 2018)* (pp. 5747–5754).

Cai, X., Zheng, Y., Yang, L., Dai, T., & Guo, L. (2019). Bibliographic network representation based personalized citation recommendation. *IEEE Access*, *7*, 457–467, doi: 10.1109/access.2018.2885507.

Chang, L. L. H., Phoa, F. K. H., & Nakano, J. (2019). A new metric for the analysis of the scientific article citation network. *IEEE Access*, *7*, 132027–132032, doi:10.1109/access.2019.2937220.

Chen, T., & Sun, Y. (2017). Task-guided and path-augmented heterogeneous network embedding for author identification. In M. de Rijke, M. Shokouhi, A. Tomkins, & M. Zhang (Eds.), *Proceedings of the 10th ACM international conference on web search and data mining (WSDM 2017)* (pp. 295–304), doi:10.1145/3018661.

Dong, Y., Chawla, N. V., & Swami, A. (2017). metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2017)* (pp. 135–144), doi:10.1145/3097983.3098036.

Du, B., & Tong, H. (2019). MrMine: Multi-resolution multi-network embedding. In W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, & J. X. Yu (Eds.), *Proceedings of the 28th ACM international conference on information and knowledge management (CIKM 2019)* (pp. 479–488), doi:10.1145/3357384.3357944.

Ferrer-Sapena, A., & Pérez, E. A. S. (2019). Inter-field nonlinear transformation of journal impact indicators: The case of the h-index. *Journal of Interdisciplinary Mathematics*, *22*, 177–199, doi:10.1080/09720502.2019.1616913.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*, 35, doi:10.2307/3033543.

yang Fu, T., Lee, W. C., & Lei, Z. (2017). HIN2vec: Explore meta-paths in heterogeneous information networks for representation learning. In E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, & C. Li (Eds.), *Proceedings of the 26th ACM on conference on information and knowledge management (CIKM 2017)* (pp. 1797–1806), doi:10.1145/3132847.3132953.

Galam, S. (2011). Tailor based allocations for multiple authorship: A fractional gh-index. *Scientometrics*, *89*, 365–379, doi:10.1007/s11192-011-0447-1.

Ganguly, S., & Pudi, V. (2017). Paper2vec: Combining graph and text information for scientific paper representation. In J. M. Jose, C. Hauff, I. S. Altingövde, D. Song, D. Albakour, S. N. K. Watt, & J. Tait (Eds.), *Advances in information retrieval – proceedings of the 39th European conference on information retrieval (ECIR 2017)* (pp. 383–395). Aberdeen, UK: Springer, doi:10.1007/978-3-319-56608-5_30.

Gao, C., Wang, Z., Li, X., Zhang, Z., & Zeng, W. (2016). PR-index: Using the h-index and PageRank for determining true impact. *PLOS ONE*, *11*, e0161755, doi:10.1371/journal.pone.0161755.

Garfield, E. (2006). The history and meaning of the journal impact factor. *The Journal of the American Medical Association*, *295*, 90–93, doi:10.1001/jama.295.1.90.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2016)* (pp. 855–864), doi:10.1145/2939672.2939754.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*, 16569–16572, doi:10.1073/pnas.0507655102.

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification. In *Proceedings of the 29th IEEE conference on computer vision and pattern recognition (CVPR 2016)* (pp. 5375–5384).

Ganguly, J. G., Gupta, S., Varma, M., & Pudi, V. V. (2016). Author2vec: Learning author representations by combining content and link information. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, & B. Y. Zhao (Eds.), *Proceedings of the 25th international conference on world wide web (WWW 2016)* (pp. 49–50), doi:10.1145/2872518.2889382.

Jamil, S., Khan, A., Halim, Z., & Baig, A. R. (2011). Weighted MUSE for frequent sub-graph pattern finding in uncertain DBLP data. In *Proceedings of the 2nd international conference on Internet technology and applications (iTAP 2011)* , doi:10.1109/itap.2011.6006415.

Jeon, H. J., Lee, O. J., & Jung, J. J. (2019). Is performance of scholars correlated to their research collaboration patterns? *Frontiers in Big Data*, 2, doi:10.3389/fdata.2019.00039.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*, 3573–3587, doi:10.1109/TNNLS.2017.2732482.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In E. P. Xing, & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning (ICML 2014)* (pp. 1188–1196).

Lee, O. J., & Jung, J. J. (2019). Character network embedding-based plot structure discovery in narrative multimedia. In R. Akerkar, & J. J. Jung (Eds.), *Proceedings of the 9th international conference on web intelligence, mining and semantics (WIMS 2019)* , doi:10.1145/3326467.3326485.

Lee, O. J., & Jung, J. J. (2020a]). Story embedding: Learning distributed representations of stories based on character networks. *Artificial Intelligence*, *281*, 103235. https://doi.org/10.1016/j.artint.2020.103235

Lee, O. J., & Jung, J. J. (2020b]). Story embedding: Learning distributed representations of stories based on character networks (extended abstract). In C. Bessiere (Ed.), *Proceedings of the 29th international joint conference on artificial intelligence (IJCAI 2020) and the 17th Pacific Rim international conference on artificial intelligence (PRICAI 2020)* (pp. 5070–5074), doi:10.24963/ijcai.2020/709.

Lee, O. J., Jung, J. J., & Kim, J. T. (2020). Learning hierarchical representations of stories by using multi-layered structures in narrative multimedia. *Sensors*, *20*, 1978, doi:10.3390/s20071978.

Li, T., Zhang, J., Yu, P. S., Zhang, Y., & Yan, Y. (2018). Deep dynamic network embedding for link prediction. *IEEE Access*, *6*, 29219–29230, doi:10.1109/access.2018.2839770.

Lippi, G., & Mattiuzzi, C. (2017). Scientist impact factor (SIF): A new metric for improving scientists' evaluation? *Annals of Translational Medicine*, *5*, 303–303, doi:10.21037/atm.2017.06.24.

Loudcher, S., Jakawat, W., Morales, E. P. S., & Favre, C. (2015). Combining OLAP and information networks for bibliographic data analysis: A survey. *Scientometrics*, *103*, 471–487, doi:10.1007/s11192-015-1539-0.

Ma, X., Wang, R., & Zhang, Y. (2019). Author name disambiguation in heterogeneous academic networks. In W. Ni, X. Wang, W. Song, & Y. Li (Eds.), *Web information systems and applications – Proceedings of the 16th international conference on web information systems and applications (WISA 2019)* (pp. 126–137), doi:10.1007/978-3-030-30952-7_15.

Mariani, M. S., Medo, M., & Zhang, Y. C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, *10*, 1207–1223. https://doi.org/10.1016/j.joi.2016.10.005

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26: Proceedings of 27th annual conference on neural information processing systems (NIPS 2013)* (pp. 3111–3119).

Narayanan, A., Chandramohan, M., Chen, L., Liu, Y., & Saminathan, S. (2016). *subgraph2vec: Learning distributed representations of rooted sub-graphs from large graphs.* , arXiv preprint: 1606.08928.

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., & Jaiswal, S. (2017). *graph2vec: Learning distributed representations of graphs.* Computing Research Repository (CoRR) abs/1707.05005.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66.* Stanford InfoLab.

Perianes-Rodríguez, A., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Gómez, C. O., & Moya-Anegón, F. (2009). Synthetic hybrid indicators based on scientific collaboration to quantify and evaluate individual research results. *Journal of Informetrics*, *3*, 91–101. https://doi.org/10.1016/j.joi.2008.12.001

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, & R. Ghani (Eds.), *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2014)* (pp. 701–710), doi:10.1145/2623330.2623732.

Reyes-Gonzalez, L., Gonzalez-Brambila, C. N., & Veloso, F. (2016). Using co-authorship and citation analysis to identify research groups: A new way to assess performance. *Scientometrics*, *108*, 1171–1191, doi:10.1007/s11192-016-2029-8.

Ribeiro, L. F., Saverese, P. H., & Figueiredo, D. R. (2017). struc2vec. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2017)* (pp. 385–394), doi:10.1145/3097983.3098061.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*, 581–603, doi:10.1007/bf02289527.

Shen, Q., Wu, T., Yang, H., Wu, Y., Qu, H., & Cui, W. (2017). NameClarifier: A visual analytics system for author name disambiguation. *IEEE Transactions on Visualization and Computer Graphics*, *23*, 141–150, doi:10.1109/tvcg.2016.2598465.

Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., & Borgwardt, K. M. (2011). Weisfeiler–Lehman graph kernels. *Journal of Machine Learning Research*, *12*, 2539–2561.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. P., & Wang, K. (2015). An overview of microsoft academic service (MAS) and applications. In A. Gangemi, S. Leonardi, & A. Panconesi (Eds.), *Proceedings of the 24th international conference on world wide web (WWW 2015)* (pp. 243–246), doi:10.1145/2740908.2742839.

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of the 8th IEEE international conference on data mining (ICDM 2008)* (pp. 1055–1060), doi:10.1109/icdm.2008.71.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale information network embedding. In A. Gangemi, S. Leonardi, & A. Panconesi (Eds.), *Proceedings of the 24th international conference on world wide web (WWW 2015)* (pp. 1067–1077), doi:10.1145/2736277.2741093.

Vaidya, J. S. (2005). V-index: A fairer index to quantify an individual's research output capacity. *The BMJ*, *331*, 13394–21340, doi:10.1136/bmj.331.7528.1339-c.

Waheed, W., Imran, M., Raza, B., Malik, A. K., & Khattak, H. A. (2019). A hybrid approach toward research paper recommendation using centrality measures and author ranking. *IEEE Access*, *7*, 33145–33158, doi:10.1109/access.2019.2900520.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*, 365–391, doi:10.1016/j.joi.2016.02.007.

Wang, D., Cui, P., & Zhu, W. (2016). Structural deep network embedding. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2016)* (pp. 1225–1234), doi:10.1145/2939672.2939753.

Wu, Z., Lin, Y., Wang, J., & Gregory, S. (2016). Link prediction with node clustering coefficient. *Physica A: Statistical Mechanics and its Applications*, *452*, 1–8, doi:10.1016/j.physa.2016.01.038.

Xia, F., Chen, Z., Wang, W., Li, J., & Yang, L. T. (2014). MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *IEEE Transactions on Emerging Topics in Computing*, *2*, 364–375, doi:10.1109/tetc.2014.2356505.

Xiao, C., Han, J., Fan, W., Wang, S., Huang, R., & Zhang, Y. (2019). Predicting scientific impact via heterogeneous academic network embedding. In A. C. Nayak, & A. Sharma (Eds.), *Trends in artificial intelligence – Proceedings of the 16th Pacific Rim international conference on artificial intelligence (PRICAI 2019)* (pp. 555–568), doi:10.1007/978-3-030-29911-8_43.

Yan, X., Zhai, L., & Fan, W. (2013). C-index: A weighted network node centrality measure for collaboration competence. *Journal of Informetrics*, *7*, 223–239. https://doi.org/10.1016/j.joi.2012.11.004

Yanardag, P., & Vishwanathan, S. (2015). Deep graph kernels. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, & G. Williams (Eds.), *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2015)* (pp. 1365–1374), doi:10.1145/2783258.2783417.

Ye, Q., Li, T., & Law, R. (2011). A coauthorship network analysis of tourism and hospitality research collaboration. *Journal of Hospitality & Tourism Research*, *37*, 51–76, doi:10.1177/1096348011425500.

Yu, S., Bedru, H. D., Lee, I., & Xia, F. (2019). Science of scientific team science: A survey. *Computer Science Review*, *31*, 72–83. https://doi.org/10.1016/j.cosrev.2018.12.001

Zhang, B., & Hasan, M. A. (2017). Name disambiguation in anonymized graphs using network embedding. In E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, & C. Li (Eds.), *Proceedings of the 2017 ACM on conference on information and knowledge management (CIKM 2017)* (pp. 1239–1248), doi:10.1145/3132847.3132873.

Zhang, S., Zhao, D., Cheng, R., Cheng, J., & Wang, H. (2016). Finding influential papers in citation networks. In *Proceedings of the IEEE 1st international conference on data science in cyberspace (DSC 2016)* (pp. 658–662), doi:10.1109/dsc.2016.55.

Zhang, Y., Ma, J., Wang, Z., Chen, B., & Yu, Y. (2017). Collective topical PageRank: A model to evaluate the topic-dependent academic impact of scientific papers. *Scientometrics*, *114*, 1345–1372, doi:10.1007/s11192-017-2626-1.

Zhang, Y., Zhao, F., & Lu, J. (2019). P2v: Large-scale academic paper embedding. *Scientometrics*, *121*, 399–432, doi:10.1007/s11192-019-03206-9.

Zhou, D., Orshanskiy, S. A., Zha, H., & Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 7th IEEE international conference on data mining (ICDM 2007)* (pp. 739–744), doi:10.1109/icdm.2007.57.

Zhou, X., Ding, L., Li, Z., & Wan, R. (2017). Collaborator recommendation in heterogeneous bibliographic networks using random walks. *Information Retrieval Journal*, *20*, 317–337, doi:10.1007/s10791-017-9300-3.