Heterogeneous academic network embedding based multivariate random-walk model for predicting scientific impact

Chunjing Xiao1 · Leilei Sun1 · Jianing Han2 · Yongwei Qiao3

Abstract
The prediction of current scientific impact of papers and authors has been extensively studied to help researchers find valuable papers and recent research directions, also help policymakers make recruitment decisions or funding allocation. However, how to accurately evaluate the future impact of them, especially for new papers and young researchers, is the focus of scientific impact prediction research, and is less explored. Existing graph-based methods heavily depend on the global structure information of heterogeneous academic network and ignore the local structure information and text information, which may provide important clues to identify influential papers and authors with novel perspective. In this paper, we propose a hybrid model called ESMR to predict the future influence of papers and authors by mainly exploiting these information mentioned above. Specifically, we first put forward a novel network embedding-based model, which can capture not only the local structure information, but also the text information of papers into a unified embedding representation. Then, the future impact of papers and authors is mutually ranked by integrating the learned embedding representations into a multivariate random-walk model. Empirical results on two real datasets demonstrate that the proposed method significantly outperforms the existing state-of-the-art ranking methods.

Keywords Scientific impact · Heterogeneous academic networks · Network embedding · Multivariate random walk

1 Introduction

Accurately assessing the potential importance of papers and authors has attracted rising research attention, and became one of the centric research issues in scientometric recently. That can help researchers catch up the most recent research directions, direct policymakers in recruitment decisions, funding allocation or expert finding [1–4]. So far, most remarkable works have focused on ranking the current importance of papers and authors [2, 5, 6], and proposed some more complicated metrics, such as h-index [7] and s-index [8]. These ranking methods can be roughly divided into citation-count based methods and PageRank-based methods. Citation count is a simple but widely used measurement to evaluate the popularity of papers and authors [7, 9]. The major limitation is that such methods only consider the popularity of papers or authors, but ignore the importance of the citations themselves. To overcome this shortcoming, PageRank-based methods (i.e., univariate random walk model) are proposed to rank the authorities of papers or authors by iteratively computing the entire citation or co-author network. In PageRank-based methods, the paper prestige can be propagated through the citation relationship among the papers [10], which is a much reasonable way for literature ranking compared to citation-count based methods. However, how to accurately identify potential papers and researchers, and predict their future impact is less explored, especially for new papers and young researchers.

The graph-based models, whether univariate or multivariate random walk, are considered as the state-of-the-arts and widely used to rank and predict future impact of papers [3, 11–14]. The univariate random walk methods is firstly used to construct a single homogeneous network or split the heterogeneous academic network (HAN) into
several homogeneous ones [12]. Such methods ignore the different influences among different types of objects. To address this issue, multivariate random walk models are proposed to rank multiple objects simultaneously [3, 5]. A key assumption of these algorithms is that the authority of the papers and the reputation of their authors are mutually reinforced. However, these methods (including both univariate or multivariate random-walk algorithms) aim at capturing the global structure information in a simple way to recursively calculate the ranking scores. Here the global structure information is topological characteristics of graphs, that is usually used to compute the transition matrix of random-walk model. While the local structure information is relations among entities in graphs. The local structure information (i.e., local similarity between entities) and the text information of papers (i.e., paper topics, words included in papers, words included in topics), which is helpful to identify influential papers and authors, can not be directly taken into account [15]. Therefore, the prediction accuracy of existing models are relatively poor because of ignoring these two kinds of information.

In fact, both the two types of information are essential to predict the future impact of authors and papers, especially for new published papers and young authors. The new published papers are those have been published recently, while young authors are researchers who began to publish papers lately. Because the structure information of their citations or co-authors is sparse, that leads to lower prediction performance. For new published papers and young authors, they can establish the direct and indirect relations between them and other authors or papers through the information of topics and words contained in paper and the local structure information. Specifically, we can use the local structure information to boost the presentations of links between nodes. Similarly, the text information of papers can be used to better capture potentially research hotspot. Some researches combine various kinds of information with multivariate random-walk or other models to improve the accuracy of prediction, such as publication time [12, 16], author order factors [16], early citations [17], journal impact factor [17], text features [12, 18], topical authority [17, 19], and so on. But it is difficult to express the information well in a unified representation to integrate those into prediction approaches. Cluster-based zero-shot learning can deal with many types of data by working on vector space [20]. Recent advances based on network embedding [21–23] have been extensively studied to learn a unified low-dimensional representation for different kinds of entities in the heterogeneous network. Motivated by these, we propose a network embedding based model to gain better prediction, which can simultaneously take the local structural and the text information into account.

In this paper, we propose a novel model called ESMR, which adds the learned Embeddings and global Structure information to Multivariate Random-walk, to predict the future scientific influence of papers and authors. More specifically, a heterogeneous academic network embedding model is first designed to learn the local structural and text information simultaneously. Then the future scientific impact of papers and authors is comprehensively predicted by using the multivariate random walk algorithm, whose transition matrix is computed by combining the learned embeddings with global structural information. Extensive experiments on two datasets have been conducted and the experimental results demonstrate that the performance of ESMR is significantly better than the existing state-of-the-art methods.

We summarize our main contributions as follows:

- We design a heterogeneous academic network (HAN), which includes multiple connections between different kinds of entities, especially for the connections between authors and research topics or words of their papers. This does help to predict scientific impact of new papers and young authors.
- A network embedding model is proposed to learn a unified representation for different types of nodes, which makes the ranking process more easier by using a multivariate random walk.
- Extensive experiments on two datasets have been conducted, and the experimental results demonstrated that ESMR can accurately predict the future impact of papers and author, especially for the new papers and authors.

2 Related work

The earliest works on scientific publication ranking are citation count based methods. Although very simple, citation count is widely used to measure the importance of papers and researchers. Based on citation count, several more complicated metrics are proposed, such as $h$-index [7], $c$-index [24] and $s$-index [8]. However, all the citation-count based methods do not consider the available network structure and only focus on citation popularity.

PageRank-based methods initially have been proposed to rank papers or authors on the homogeneous networks of citation or co-author network, which propagates the paper prestige through the citation relationship among the papers [10]. Although they can give the current influence of papers or authors, it is difficult to predict the future influence of them. It needs many types of related factors when we used it to accurately predict the future influence of papers and authors. However, it includes limited information to do it on the homogeneous networks of citation or co-author network.
In fact, an academic network is heterogeneous and is composed of various different kinds of networks, such as co-author network, paper citation network, and venue-paper network. The graph-based models including univariate and multivariate random walk techniques are widely used to predict future impact of papers. These methods first construct a heterogeneous academic network. Then the univariate random walk techniques usually split the heterogeneous academic network into several homogeneous ones by treating all the nodes and edges as the same type (like PageRank-based methods mentioned above) [12]. Such straightforward methods ignore the different influences between different types of objects, and thus limit the effectiveness in ranking different kinds of objects. For example, the authorities of papers can enhance the academic reputations of their authors, conversely, the reputable authors can also increase the authorities of their papers. Thus, some multivariate random-walk techniques have been proposed to rank multiple entities simultaneously to identify the future influence of papers and researchers. The Co-Rank algorithm was the first method to improve the ranking results for both papers and researchers by using citation network, co-author network, and the social network of authors. Most of later related works followed or extended Co-Rank simultaneously rank a kind or different types of entities (such as papers, authors) [5, 11, 12].

Following these methods, various kinds of information about papers and authors are integrated into the multivariate random-walk framework to further improve the accuracy of prediction. Sayyadi et al. [11] and Wang et al. [12] applied the time information about papers to the multivariate random-walk ranking model to predict the future citations of papers under the assumption that new published papers are easier to be cited than older ones. Wu et al. [16] proposed TAORank, which considers mutual influence among scholarly entities, which includes the publication time and author order in scholarly papers. Wang et al. [12] incorporated the text information because they thought that it is useful to improve the predicting results. Chaturvedi and Snigda [18] analyzed the usefulness of text features and got the conclusion that the most accurate prediction result can be obtained from combining the metadata and text features. Dong et al. [19] and Giovanni et al. [17] found that topical authority and publication venue were crucial to these effective predictions. Liang et al. [5] integrated the multinomial multidimensional relationships between papers and authors into ranking model. However, the limitation of these methods is that the rich information is not expressed in a unified representation to adequately use it for prediction.

In recent years, network embedding-based methods have received widespread attention for their ability of learning unified low-dimensional vectors for different kinds of entities in a network while the structure information is preserved. Various network embedding algorithms have been put forward for multiple tasks, such as link prediction [26], node classification [20, 21, 23, 27], community detection [28], and recommendation task [29]. These provide a dawn for us to solve this problem. However, these methods have not been extended to scientific influence prediction task. In this paper, a novel network embedding-based method called ESMR has been developed to predict future scientific impact, which can simultaneously capture the local and global structure information and text information from HAN.

3 Methodology

The goal of our proposed ESMR is to predict future impact of papers and authors by integrating learned embedding representations of entities in HAN into multivariate random-walk model, which can capture the local structure information of HAN and the rich text information related to the papers. The process of ESMR is shown as Fig. 1, which consists of three parts: (1) constructing the heterogeneous academic network (HAN), which includes the paper citation network (PCN), text information network (TIN), paper-author network (PAN) and co-author network (CAN); (2) Learning a unified embedding representations for all entities of the HAN, in which the local structure information and text information are preserved; (3) predicting the future impact of papers and authors by integrating these learned representations into a multivariate random walk based Co-Ranking model.

3.1 Heterogeneous academic network definition

In this subsection, different types of networks are defined and can be considered to be the formal construction of the heterogeneous academic network.

Definition 1 Paper Citation Network (PCN). Paper citation network is denoted as $G_{pp} = (P, E_{pp}, F_{pp})$, where $P$ is a set of papers, $E_{pp}$ is the directed edges representing the citation relationships among the papers. $F_{pp}$ is a set of edge weights. $F_{pp} = \{ f_{ij} : p_i \rightarrow p_j \rightarrow R^+ | e_{ij} \in E_{pp} \}$.

That is to say if paper $p_i$ cites $p_j$, there exists an edge $e_{ij} \in E_{pp}$ from $p_i$ to $p_j$, whose weight is represented by $f_{ij} \in F_{pp}$. It relates to the time when paper $p_i$ cites $p_j$ and the time span between the current time and their citation time, and is calculated by (3). So the paper citation network is a time-aware graph, in which an edge established newly can obtain the greater weight because the papers
frequently cited recently are much more likely to be cited by other papers. The paper citation network can capture the citation relationship among the papers. However, the newly published papers with few citations may not be well represented. Generally, the citation relationship between two papers is established because of the similar research topics or contents, so the text information of papers can be helpful to learn the better representations of papers. The words and topics in the papers are used to capture the text information, which is added to the embedding of the paper citation network.

Definition 2 Text Information Network of A Paper (TINp). Given a paper \( p_i \in P \), let \( G^i_p = (V^i_p, E^i_p, F^i_p) \) be its text information network, where \( V^i_p = \{ p_i \cup Z^i_p \cup W^i_p \} \) is the set of nodes including the words \( W^i_p \) and topics \( Z^i_p \) contained in paper \( p_i \); \( E^i_p = \{ E^i_{pz} \cup E^i_{pw} \cup E^i_{zw} \} \) is the set of edges between \( p_i \) and its topics and words, \( F^i_p \) is the set of weights of \( E^i_p \) and \( F^i_p = \{ F^i_{pz}, F^i_{pw}, F^i_{zw} \} \).

There exists edges \( e^i_{pz} \in E^i_{pz} \) and \( e^i_{pw} \in E^i_{pw} \) if the topic \( z_i \) and words \( w_j \) are included in \( p_i \). \( e^i_{zw} \in E^i_{zw} \) if the topic \( z_j \) includes the word \( w_k \) in the paper \( p_i \). \( F^i_p \) is the set of weights of edges \( E^i_{pz}, E^i_{pw}, \) and \( E^i_{zw} \), respectively. The weight \( f^i_{pz} \in F^i_{pz} \) is calculated by LDA, while the weight \( f^i_{zw} \in F^i_{zw} \) by TF-IDF, which is described detailly in next Section. TINp of a paper \( p_i \) contains paper-topic network \( G^i_{pz} \), paper-word network \( G^i_{pw} \) and topic-word network \( G^i_{zw} \). For the published papers (especially for new papers) with only fewer citations, the topics and words information of \( p_i \) may provide us clues to represent the context of it. That will help us address the problem of less available information of them when ranking their future influences. After defining PCN and TINp, We can further define the paper citation network with text information (PCNT). The citation relationships between two papers are established largely because of the similar research topics, and these similarities can be further reflected by text content of articles. So PCN and TIN are combined into a unified network PCNT.

Definition 3 Paper citation network with text information (PCNT) [15]. Let \( G_{pp} \) and \( G^i_p \) be graphs as in Definition 1 and Definition 2, then the PCNT is defined as \( G_p = (V_p, E_p, F_p) \), where \( V_p = \{ P \cup Z_p \cup W \} \), \( E_p = \{ E_{pp} \cup E_{pz} \cup E_{pw} \cup E_{zw} \} \), and \( F_p = \{ F_{pp} \cup F_{pz} \cup F_{pw} \cup F_{zw} \} \). \( P, Z, W \) respectively represent the sets of papers, their topics and words, \( E_{pp}, E_{pz} \) and \( E_{pw} \) are the sets of edges between papers and their references, topics, words, \( E_{zw} \) is the set of edges between topics and words, \( F_{pp}, F_{pz}, F_{pw}, F_{zw} \) are the corresponding weights of the edges \( E_{pp}, E_{pz}, E_{pw}, E_{zw} \), respectively.

The PCNT can help us to accurately predict the scientific impact of new papers with few citations by effectively calculating the text semantic similarities between papers. In order to discover the relationships of authors, papers and their authors, we define the co-author network and paper-author network, respectively.

Definition 4 Co-author network (CAN). The CAN is denoted as \( G_a = (V_a, E_a, F_a) \), where \( V_a \) is a set of authors, \( E_a \) is the set of undirected edges representing collaborations among authors, \( F_a \) is the set of weights of edges \( E_a \), which is calculated by (5).
Heterogeneous academic network embedding...

It is also a time-aware weighted network, and when we calculate the weight between author \( a_i \) and \( a_j \), the number of papers coauthored by them and the time span between the current time and the time they coauthor every paper are taken into account. It assumes that an author is more likely to coauthor new papers with authors who recently coauthors with him.

Definition 5 Paper-author network (PAN). The PAN is defined as \( G_{pa} = (P \cup V_a, E_{pa}, F_{pa}) \), where \( E_{pa} \) is the set of edges between papers and authors, which connects the papers and the corresponding authors, and \( F_{pa} \) is the set of weights of edges \( E_{pa} \). There exists an edge \( e_{pi}^{ij} \in E_{pa} \) between \( p_i \) and \( a_j \) if \( a_j \) is an author of \( p_i \), and the weight \( f_{pi}^{ij} \in F_{pa} \) relates to the signature order of author \( a_j \) in paper \( p_i \), which is described detailly in next Section.

Multiple networks mentioned above can be further merged into a unified network, in which multiple connections between different types of entities are established, especially for the connections between authors and research topics or words of papers. This can help us predict scientific impact of new papers and new authors.

Definition 6 Heterogeneous academic network (HAN) [15]. Let \( G_p \) be graph as in Definition 3, \( G_a \) in Definition 4, \( G_{pa} \) in Definition 5, then the HAN is defined as \( G = (V_p \cup V_a, E, F) \), where \( E = \{E_p \cup E_a \cup E_{pa}\} \) is the set of different types of edges, and \( F = \{F_p \cup F_a \cup F_{pa}\} \) is the set of weights of the corresponding edges.

Then the heterogenous academic network can be established to learn the embeddings of entities in it. For example, there are three papers \( p_1, p_2 \) and \( p_3 \), their authors are \( a_2 \) and \( a_1, a_2 \) and \( a_3, a_1 \) and \( a_3 \), respectively. So there exists weights \( f_{12}^{pp}, f_{23}^{pp}, f_{31}^{pp} \) between \( p_1 \) and \( p_2 \), and \( p_3 \) and \( p_1, f_{21}^{aa}, f_{31}^{aa}, f_{32}^{aa} \) between \( a_2, a_1, a_2 \) and \( a_3, a_1, a_3 \), respectively. While in CAN \( f_{12}^{pp} = 1, f_{21}^{pp} = 1/2, f_{22}^{aa} = 1, f_{23}^{aa} = 1/2, f_{31}^{aa} = 1, f_{33}^{aa} = 1/2 \). If \( p_1 \) includes \( z_1, z_2 \), and \( w_2, w_3 \), there exists \( e_{1z_1}, e_{1z_2}, e_{2z_1}, e_{2z_2}, e_{3z_1}, e_{3z_2}, e_{2w_1}, e_{2w_2} \), and their weights are calculated by LDA and TF-IDF, respectively. After the representations of entities are learned, the influences of papers and authors are ranked using multivariate random-walk on HAN.

3.2 Embedding for heterogeneous academic network

3.2.1 Embedding for PCNT

As mentioned above, the PCNT \( G_p \) consists of three networks \( G_{pp}, G_{pc} \) and \( G_{pw} \), which are connected by the paper nodes. The empirical distributions of paper \( p_i \) in \( G_{pp}, G_{pc} \) and \( G_{pw} \) are uniformly expressed as \( \hat{P}(\cdot|p_i) \), and \( P(\cdot|p_i) \) are their conditional probability distributions. To learn the low-dimensional embedding representations \( p_i, z_j, w_j \) of paper \( p_i \), topic \( z_i \), and word \( w_j \), the objective is to minimize the following KL-divergence between two probability distributions.

\[
\mathcal{L}_p = - \sum_{(i,j) \in E_p} \lambda_i^p \hat{P}(v_j|p_i) \log P(v_j|p_i),
\]

where \( v_j \) is one of \( p_j, z_j \) and \( w_j \), \( \lambda_i^p \) is a unified representation, including \( \lambda_i^p, \lambda_z^p \) and \( \lambda_w^p \), which are weights of paper \( p_i \) representing the importance of \( p_i \) in \( G_{pp}, G_{pc} \) and \( G_{pw} \) and will be defined below. \( P(v_j|p_i) \) can be estimated by the following softmax function:

\[
P(v_j|p_i) = \frac{\exp(v_j^\top \cdot p_i)}{\sum_{k \in V_p} \exp(v_k^\top \cdot p_i)},
\]

where \( p_i \in \mathbb{R}^d \) and \( v_j \in \mathbb{R}^d \) are \( d \)-dimensional latent representations of \( p_i \) and \( v_j \), respectively.

\( \hat{P}(v_j|p_i) \) can be computed by \( \hat{P}(v_j|p_i) = \sum_{k \in \mathcal{R}(p_i) \omega_{ik}} \omega_{ij} \), where \( \omega_{ij} \) is the weight of edge \((p_i, v_j)\), and \( \mathcal{R}(p_i) \) is the nodes connecting to \( p_i \). As with \( \lambda_i^p \), it has different definitions in \( G_{pp}, G_{pc} \) and \( G_{pw} \).

For \( G_{pp}, v_j \in P, R(p_i) \) is the set of papers cited by \( p_i \), and \( \omega_{ij} (\omega_{ij} \in F_{pp}) \) is the edge weights between papers which represents the citation relationship. Obviously, the paper citation network is dynamic, and the citation relationships established at different years have different effects on the future influence of papers. Thus we try to capture the dynamic properties of the network by assigning different weights to the citation relations based on their set-up time. We assign higher weights to the more recent citations through the exponential decay function over time. Thus \( \omega_{ij} \) is defined as:

\[
\omega_{ij} = e^{-\rho(T_c - T_{i-j})},
\]

where \( \rho \) is a decaying parameter that has been predefined, \( T_c \) represents the current time, and \( T_{i-j} \) is the time that the citation occurs between paper \( p_i \) and \( p_j \). Furthermore, we set \( \lambda_i^p = \sum_{k\in\mathcal{C}(p_i)} \omega_{ik} \), where \( \mathcal{C}(p_i) \) is the set of papers which references \( p_i \). It represents the influence of paper \( p_i \) in the paper citation network.

For \( G_{pc}, v_j \in Z, R(p_i) \) is the set of topics that are most likely to be touched upon by \( p_i \). \( \omega_{ij} (\omega_{ij} \in F_{pc}) \) represents the likelihood that the topic \( z_j \) is included in \( p_i \) (i.e., \( \mathcal{P}(z_j|p_i) \)), which is calculated by LDA model [30]. And \( \lambda_z^p \) is defined as \( \lambda_z^p = \sum_{k\in\mathcal{R}(p_i)} \omega_{ik} \). It represents the influence of paper \( p_i \) over the topics.

For \( G_{pw}, v_j \in W, R(p_i) \) is the set of words which are included in \( p_i \). \( \omega_{ij} (\omega_{ij} \in F_{pw}) \) reflects the importance of
\( w_j \) in \( p_i \) and can be calculated by if-idf. And \( \lambda_{pi}^j \) is denoted as \( \lambda_{pi}^j = \sum_{k \in R(p_i)} a_{ik} \). It represents the influence of paper \( p_i \) over the words.

### 3.2.2 Embedding for CAN

The CAN \( G_a \) can show the influence of authors by mining the cooperation relationship among authors. We assume that the impact of two authors sharing common co-authors is similar to each other. Then similar to the PCNT, the loss function for embedding the co-author network \( G_a \) can be defined as:

\[
L_a = - \sum_{(i,j) \in E_a} \lambda_{ai}^j \hat{P}(a_j|a_i) \log P(a_j|a_i),
\]

where \( \hat{P}(a_j|a_i) = \frac{a_{ij}^a}{\sum_{k \in N(a_i)} a_{ik}^a} \). \( N(a_i) \) is the set of co-authors of author \( a_i \), and \( a_{ij}^a \) is the weight of the collaboration among co-authors. Although the number of papers co-authored by two authors reflects the closeness of their collaboration relationship, it is not fair for those young authors who do not have many co-authors. To this end, the time information is taken into account and the weight of the collaboration between author \( a_i \) and \( a_j \) is set as

\[
\omega_{ij}^a = \sum_{p_i \in C(a_i,a_j)} e^{-\rho(T_c-T_{co}^p)},
\]

where \( C(a_i,a_j) \) is the set of papers that author \( a_i \) collaborates with author \( a_j \). \( T_c \) is the current time, and \( T_{co}^p \) is the time when author \( a_i \) and \( a_j \) co-authored paper \( p_k \). \( \lambda_a^l \) represents the influence of different authors \( a_i \) in \( G_a \), and is computed by \( \lambda_{ai}^j = \sum_{j \in N(a_i)} \omega_{ij}^a \). The conditional probability \( P(a_j|a_i) \) is also calculated by the softmax function defined by (2).

### 3.2.3 Embedding for PAN

The paper-author network \( G_{pa} \) can capture the relationships between a paper and its all authors, and that should be preserved in PAN Embedding. The weight \( \omega_{ij}^{pa} \) of an edge \( (p_i, a_j) \) linking a paper \( p_i \) and its author \( a_j \) is regarded as their empirical probability which indicates the closeness between them. The joint probability \( P(p_i, a_j) \) is specified by the low-dimensional representation in the latent space. So the embedding for PAN can be learned by minimizing the following objective function.

\[
L_{pa} = - \sum_{(i,j) \in E_{pa}} \omega_{ij}^{pa} \log P(p_i, a_j),
\]

where \( \omega_{ij}^{pa} \) is set to \( \frac{1}{s} \), and \( s \) is the signature order of author \( a_i \) in paper \( p_j \). \( P(p_i, a_j) \) is defined as

\[
P(p_i, a_j) = \frac{1}{1+\exp(-p_i \cdot a_j)},
\]

where \( p_i \in \mathbb{R}^d \) and \( a_j \in \mathbb{R}^d \) are \( d \)-dimensional latent representations of \( p_i \) and \( a_j \), respectively.

### 3.2.4 Embedding for HAN

To embed the HAN by integrating all the embeddings on \( G_p, G_a \) and \( G_{pa} \), we combine the objective functions (1), (4) with (6), then jointly minimize the following objective function.

\[
L = L_p + L_a + L_{pa}.
\]

### 3.2.5 Model optimization

Direct calculation of the function in (2) [21] is both time-consuming and impractical. To address the computational challenge, negative sampling approach [31] is adopted. The probability of positive samples is maximized while the probability of negative samples is minimized as far as possible. Therefore, the objective function \( L \) can be expressed as the following formula, which uses L2-norm to avoid over-fitting and ignores some constraints.

\[
L = - \sum_{(i,j) \in E'} \lambda_{ij}^p \hat{P}(v_j|v_i) \log(\sigma(v_j^\top \cdot v_i))
- \sum_{(i,j) \notin E'} \lambda_{ij}^p \log(\sigma(-v_j^\top \cdot v_i))
- \sum_{(i,j) \in E_{pa}} \omega_{ij}^{pa} \log(\sigma(p_i^\top - a_j))
+ \lambda \sum_{n=1}^{p} ||p_n||_2 + \beta \sum_{n=1}^{a} ||a_n||_2,
\]

where \( \sigma(x) = \frac{1}{1+\exp(-x)} \) is the sigmoid function. \( \lambda, \beta \in \mathbb{R} \) are regularization coefficients. The positive samples are modeled by the first term in (8), and negative samples by the second term. Where \( E' = \{ E_{pp} \cup E_{pz} \cup E_{pw} \cup E_a \} \), \( (i, j) \notin E' \) represents a set of randomly sampled edges between \( v_i \) and \( v_j \), which are not actually included in HAN.

In order to optimize the loss function (7), the gradients for \( p_i \) and \( a_j \) can be computed by using the stochastic gradient descent algorithm.

\[
\frac{\partial L}{\partial p_i} = - \sum_{(i,j) \in E_p} \lambda_{ij}^p \hat{P}(v_j|p_i) \exp(-v_j^\top p_i) v_j
+ \sum_{(i,j) \notin E_p} \lambda_{ij}^p \frac{\exp(-v_j^\top p_i) v_j}{1+\exp(-v_j^\top p_i)} v_j
- \sum_{(i,j) \in E_{pa}} \omega_{ij}^{pa} \frac{\exp(-p_i^\top a_j)}{1+\exp(-p_i^\top a_j)} a_j + \lambda \sum_{r=1}^{d_p} 2(p_i^r),
\]

where \( \hat{P}(v_j|p_i) = \frac{1}{1+\exp(-p_i^\top a_j)} \).
\[ \frac{\partial L}{\partial a_i} = - \sum_{(i,j) \in E_a} \lambda_{ij} \frac{\hat{p}(a_j | a_i) \exp(-a_j^\top a_i)}{1 + \exp(-a_j^\top a_i)} a_j \]
\[ + \sum_{(i,j) \not\in E_a} \lambda_{ij} \frac{\hat{p}(a_i | a_j) \exp(-a_i^\top a_j)}{1 + \exp(-a_i^\top a_j)} a_j \]
\[ - \sum_{(j,i) \in E_p} \alpha_{pa} \frac{\exp(-p_j^\top a_i)}{1 + \exp(-p_j^\top a_i)} p_j + \beta \sum_{r=1}^{d_p} 2(a_i^r), \tag{10} \]

where \(d_p\) and \(d_a\) respectively represent the dimension of vectors \(p\) and \(a\), and here \(d_p = d_a = d\). We do not detailedly show the gradients for \(z_i\) and \(w_i\), which can be derived in the similar way.

For each iteration, we adopt the backtracking line search [22] to obtain the most suitable learning rate. The complexity of Algorithm 1 is proportional to the complexity of the gradients of vertex embeddings. Let \(n\) be the number of pairs of vertices with edges, \(k\) be the iteration times, and \(d_p\) and \(d_a\) are the dimensions of \(v_p\) and \(v_a\), respectively. Then its complexity is \(O(n d_p \times d_a k)\). Therefore, it is easy to see that the training can be done in polynomial time. The detailed training process has shown in Algorithm 1 [15].

**Algorithm 1** The training process for embedding HAN.

1: **Input:** \(G(V, E, F)\), learning rate \(\eta\), dimensions of vectors \(d\), negative sampling rate \(k\), regularized coefficients \(\lambda\) and \(\beta\)
2: **Output:** latent vectors of papers, topics, words and authors
3: initializing vectors of \(p, z, w\) and \(a\);
4: **while** (not converged) **do**
5: for all \(P\), calculating \(\hat{p}\);
6: setting up \(\eta_i\) by using backtracking line search;
7: for all \(P\), updating \(p\);
8: for all \(P\) and \(Z\), updating \(p\) and \(z\);
9: for all \(P\) and \(W\), updating \(p\) and \(w\);
10: **end while**
11: for all \(V_a\), pre-training \(a\);
12: **while** (not converged) **do**
13: optimizing \(V_p(p, z, w)\) based on (9) and \(\frac{\partial L}{\partial z}, \frac{\partial L}{\partial w}\) when fix \(V_p\);
14: similarly optimizing \(V_a(a)\) based on (10) when fix \(V_p\);
15: **end while**
16: **return** latent vectors of \(p, z, w\) and \(a\)

### 3.3 Predicting the future scientific impact of papers and authors

In this section, we will introduce how does ESMR predict the future influence of papers and authors. By integrating all the available information through HAN embedding, different entities with similar potential influence are considered closer to each other in the learned latent representation space. Then based on the learned entities embedding, the cosine similarity is used to measure the similarity between them. For example, the similarity between two papers can be calculated by \(Sim(p_i, p_j) = \frac{p_i \cdot p_j}{\|p_i\| \cdot \|p_j\|}\). Thus, to be used in the multivariate random-walk model, the transition matrix of PCN can be represented as:

\[ M_{pp}(j, i) = \begin{cases} 
\gamma \frac{Sim(p_i, p_j)}{\sum_{p_r \in N(p_i)} Sim(p_i, p_r)} + (1 - \gamma) \frac{\lambda_{ip}}{deg(p_i)}, & e_{pp} \in E_{pp} \\
0, & \text{otherwise}
\end{cases} \]  \(\tag{11}\)

where \(N(p_i)\) and \(deg(p_i)\) respectively represent the sets of neighbors and out-degree nodes of \(p_i\), \(\gamma\) is used as a adjustable parameter to balance factors affecting the transition probability. In a similar way, \(M_{aa}, M_{pa}, M_{ap}\) can be easily learned.

At last, the intra and inter-network multivariate random-walk on HAN uses these transition matrices to calculate the future influence of papers and authors. Each iteration process is defined as the following equations:

\[ p^{(t+1)} = \alpha_{pp} M_{pp} p^{(t)} + \beta_{pa} M_{pa} a^{(t)}, \]
\[ a^{(t+1)} = \alpha_{aa} M_{aa} a^{(t)} + \beta_{ap} M_{ap} p^{(t)}, \]

where \(p^{(t)}\) and \(a^{(t)}\) are predicted score vectors at time \(t\). \(\alpha_{pp}\) and \(\beta_{pa}\) are influence weights of other papers and authors on one specific paper, while \(\alpha_{aa}\) and \(\beta_{ap}\) are on one specific author. Thus, the stationary vectors can be obtained by iterating the (12) and (13) until convergence.

For example, there are papers \(p_1, p_2, p_3\) and their authors \(a_1, a_2, a_3\) in HAN mentioned above, and their embeddings have been learned by using network embedding. Then we can predict their score vectors \(v_{p_1}, v_{p_2}, v_{p_3}\) and \(v_{a_1}, v_{a_2}, v_{a_3}\) using multivariate random-walk based on calculated transition matrices \(M_{pp}, M_{aa}, M_{pa}\) and \(M_{ap}\). We respectively rank the scores vectors \(v_{p_1}, v_{p_2}, v_{p_3}\) and \(v_{a_1}, v_{a_2}, v_{a_3}\) in descending order. The rankings of their influences are \(p_1 > p_3 > p_2\) and \(a_2 > a_3 > a_1\) if the rankings of their score vectors are \(v_{p_1} > v_{p_3} > v_{p_2}\) and \(v_{a_2} > v_{a_3} > v_{a_1}\).

### 4 Experimental results

#### 4.1 Datasets

ESMR is evaluated by using the following two public datasets. One is the ACL Anthology Network (AAN), AAN is the complete collection of computational linguistics papers published by ACL. It contains 23,766 papers.
published before 2014, 18,862 authors, and 124,857 citations among these papers.

Another dataset is the Academic Social Network of AMiner Dataset, which is released by AMiner. We use two versions called AMiner2014\(^1\) and AMiner2020.\(^2\) AMiner2014 includes 2,092,356 papers published before 2014 and their 8,024,869 citations. AMiner2020 includes 4,894,081 papers published before 2020 and 45,564,149 citations. The metadata for each paper contains the following information: paper ID, paper title, author list, author affiliation, published year, abstract and the list of references.

Firstly, the datasets are preprocessed as follows. Papers published after 1998 in AMiner2014 and 2000 in AMiner2020 were selected for evaluating the predicting performance. Then the papers without sufficient metadata, such as without author, publication time, reference or citation, are removed, because the impact of such papers is hard to evaluate. Then, authors are extracted from these selected papers and their effects are predicted. Finally, we obtain 19,564 papers, 91,498 citations in the AAN dataset. In the AMiner dataset2014, we obtain 328,971 papers and 2,732,340 citations, while in AMiner2020, 809,392 papers and 10,222,111 citations are selected.

### 4.2 Experiment setup

#### 4.2.1 Ground truth

Owing to the lack of criteria, it is a challenge to evaluate the performance of almost all works. We use the number of future citations as the ground truth to evaluate it, which is adopted by recent works [3, 12].

The dataset is divided into two parts according to a historical time point: the training part and the test part. The training part is used to obtain the estimated ranking lists of papers and authors published before the historical time point. The test part is used to calculate the future citation number for each paper, and then the ground truth lists of papers and authors can be obtained by ranking them in the descending order. Finally, the results is reported by comparing the similarities of the two ranking lists.

In this paper, the papers are divided into the training and test set based on whether or not they published before 2009 in AAN and AMiner2014, before 2015 in AMiner2020. To get better ground truth, the future citations of papers are calculated in different periods. For example, the future citations of papers published before 2009 are calculated in periods 2010, 2010-2012, and 2010-2014 in AMiner2014. The results obviously demonstrate that it is not fair to the new papers if using the results in 2010. None of top-10 papers of 2010 was published after 2005. While the impact rankings of these papers is highly consistent in the next three years and the next five years. 9 out of the top-10 papers of 2010-2014 are in the top-10 rankings of 2010-2012, and 47 out of the top-50 papers of 2010-2014 are in the top-50 rankings of 2010-2012. Thus, the prediction results between three and five years are good and stable across time. The citations obtained in next five years are regarded as the ground truth to evaluate our prediction results.

#### 4.2.2 Evaluation metrics

There are two widely used metrics to evaluate the performance. One is recommendation intensity \(RI\) [12]. The intuition behind the \(RI\) is that, given two ranking lists \(R1\) and \(R2\) with the top-\(k\) results, \(R1\) is better than \(R2\) if \(R1\) returns more objects matching the ground truth ranking list, and the matched objects are at the front of the top-\(k\) list. Assuming \(R\) is the top-\(k\) returned objects of a ranking approach and \(L\) is the list of ground truth, for each object \(P_i\) in \(R\) with the ranked order \(o_i\), the recommendation intensity of \(P_i\) at \(k\) can be defined as

\[
RI(P_i)@k = \begin{cases} 
1 + (k - o_i)/k, & P_i \in L \\
0, & P_i \notin L
\end{cases}
\]  

(14)

Based on each object’s recommendation intensity, the recommendation intensity of the top-\(k\) list \(R\) can be defined as \(RI(R)@k = \sum_{p_i \in R} RI(P_i)@k\). As mentioned in [12], \(RI\) will degenerate to precision when taking the top-\(k\) list \(R\) as un-ordered and dividing \(RI(R)@k\) by \(k\).

The other is Normalized Discounted Cumulative Gain (NDCG), which is commonly used in sorting algorithms. The factors considered in NDCG are the relevance of the ranking lists and the sorting position. Its intuition is to divide the relevance of each ranking list into multiple levels for scoring. The higher the level is, the higher the importance is. Then considering the position information of each ranking list, the higher the order position is, the higher the importance is. It can be calculated as follows.

\[
NDCG@k = \frac{DCG@k}{IDCG@k} = \frac{\sum_{i=1}^{k} 2^{rel_i} - 1}{\sum_{i=1}^{\|REL\|} 2^{rel_i} - 1} \frac{\log_{2}(i+1)}{\log_{2}(i+1)}
\]  

(15)

where \(rel_i\) represents the correlation of the \(i\)-th result, and \(\|REL\|\) represents the set of Top-\(k\) results selected after sorting the correlation in descending order. Obviously, the higher the value of NDCG is, the better the ranking result is.

### 4.3 Baselines

To evaluate ESMR, the following methods are used to compare with it.
• **MRCoRank (MR).** MRCoRank is the state-of-the-art graph-based method to rank the future impact of papers, authors, and venues simultaneously. It takes time, text and structure information into account when using a mutual reinforcement framework to predict the results [12].

• **FutureRank (FR).** FutureRank is a representative model to rank the future impact of papers by fusing the relevant information related to papers (like authors, citations, and publication time) [11].

• **PageRank (PR).** PageRank is a base model for many graph-based ranking methods [32], which is used to compare with ESMR having the same weights of edges.

• **LINE+CoRank (LCR.)** LCR is a method that the paper and author embeddings are learned by using LINE [21]. Then the learned embeddings are combined with our ranking algorithm described in Section 3.3.

• **EOE+CoRank (ECR).** ECR is an advanced method that integrates the paper and author embeddings learned by EOE [22] into our proposed ranking model described in Section 3.3.

In addition, ESMR has the two different variations: ESMR without Text information (ESMR-T) which shows the effectiveness of the text information, and ESMR without the network embedding model (ESMR-NE) which studies the necessity of embedding process for improving the prediction performance.

### 4.4 Parameter sensitivity analysis

There are several parameters to be learned in our model. The initial values of each dimension of the vectors are random, which appears in the vectors of authors, papers, topics and words. All of them are randomly generated between 0 and 1, which takes the current system time as seed. Other parameters are learned by experiments. We evaluate the performance of our model by varying these parameters in a predefined and default search range to determine the values of them. In the process of training ESMR, the negative sampling rate is set to 5. The regularization coefficients \( \lambda, \beta \) are specified as 1. To study the influence of dimensions, the dimensions of papers and authors are varied from 20 to 200. The results of top-20 paper ranking and author ranking are shown in Fig. 2a–d. The result shows that the performance of ESMR slightly varies in different dimensions and it is reasonable to select 100 as their dimensional values by balancing computational complexity and algorithm performance. So, 100 is used as the default dimension setting in the following experiments.

Furthermore, here are four parameters in the process of ranking, \( \alpha_{pp}, \alpha_{aa}, \beta_{pa} \) and \( \beta_{ap} \). Take the AAN dataset as an example, Fig. 3a, b shows the effect of \( \alpha_{pp} \) on papers, and Fig. 3c, d shows the effect of \( \alpha_{aa} \) on authors. For \( \alpha_{pp} \), it is set to 0.3 for getting the best result. For \( \alpha_{aa} \), 0.15 is a reasonable choice. Whether it is too large or too small for both \( \alpha_{pp} \) and \( \alpha_{aa} \) will reduce the performance. In the following experiments, for the AAN dataset, \( \alpha_{pp} = 0.3, \alpha_{aa} = 0.15, \beta_{pa} = 0.3 \) and \( \beta_{ap} = 0.85 \) are the default parameter settings. And for the two AMiner datasets, when \( \alpha_{pp} = 0.6, \alpha_{aa} = 0.2, \beta_{pa} = 0.35 \) and \( \beta_{ap} = 0.8 \), the performance of ESMR is best one. For AAN, 0.3 is the value of parameter \( \gamma \) in (11) while for the two AMiner datasets, the value is 0.6.

The parameters of baselines are same as the settings in the corresponding papers. For MRCoRank, \( \alpha_{pp} = 0.6, \alpha_{aa} = 0.5, \beta_{pa} = 0.2, \alpha_f = \alpha_v = 0.5 \) and \( \gamma_v = \gamma_{vp} = \gamma_{ia} = 0.4 \). For FutureRank, \( \alpha = 0.19, \beta = 0.02 \) and \( \gamma = 0.79 \). For PageRank, \( \alpha = 0.85 \), and random jump with a probability of 0.15. For LCR and ECR, Vector dimension is 100, other CoRank parameters are same as our proposed method.

### 4.5 Ranking results of paper impact

The performance of ESMR is quantitatively compared with all baselines, the ranking results of papers on the AAN dataset are illustrated in Fig. 4.

Figure 4 demonstrates that the performance of ESMR is better than all baselines with different k. ECR and ESMR-T perform better than most of other baselines for both NDCG@k and RI@k, but it is still inferior to ESMR. For NDCG@50 and RI@50, ESMR outperforms ESMR-T by 21% and 27%, ECR by 20% and 25%, improves over MRCoRank which achieves the best performance in all other baselines by 33% and 42% on two metrics. A possible explanation is that ESMR-T and ECR fail to capture the text information of papers, MRCoRank does not use network embedding, whereas ESMR adds all of them to improve the performance. ESMR-T is consistent with ECR in most cases, sometimes better than ECR. The possible reason is that they integrate network embedding with corank. The performance of LCR is lower than that of ESMR-T and ECR. This is because LCR is a graph based method built on random walk, and it fails to capture all the relations between entities. But LCR performs better than MRCoRank for RI@k, and EMSR-NE is inferior to all methods except FutureRank and PageRank. All these verify network embedding indeed facilitate to generate effective impact prediction.

Next, we turn to the experiments on the Aminer dataset and the papers published in the same year and in the same research community are selected to evaluate the prediction results. It is based on the following two considerations: (1) After the top-100 papers in the ground truth are listed, we discover that the ground truth obviously tends to the older ones. For example, the number of papers published before
2006 account for more than 80% in AMiner2014. (2) The difference in the number of future citations is very large for the papers published in various kinds of research fields. For example, in AMiner2014 dataset, the future citations of the most cited papers in the field of Information Security is 245, while there are only 37 papers in Artificial Intelligence obtaining more than 245 citations. Therefore, for giving a fair comparison, the results are evaluated in term of different publication years and research fields, respectively.

The results in fields of Artificial Intelligence (AI) and Database(DB) in different years are selected to evaluate the paper ranking performance and NDCG@20 and RI@20 are taken as metrics, respectively.

The results on AMiner2014 dataset in 2001, 2003, 2005 and 2007 are shown in Fig. 5. The results are basically consistent with them on AAN, but there are still some differences. For the ranking results in AI, as shown in Fig. 5a, b, ESMR generally outperforms LCR and ECR, which is better than MRCoRank and FutureRank. ESMR-T does well in most cases, and its performance is competitive with LCR. For the results on NDCG@20 and RI@20, ESMR respectively performs better than ECR by 9.7% and 10.2% in 2001, than LCR by 9.8% and 14.6% in 2003, 17.5% and 21.2% in 2005, 21% and 29% in 2007. That confirms that ESMR is effective to rank the future influence of papers. ESMR generally outperforms ESMR-T, that proves the text information of papers is useful to improve the accuracy.

For the prediction results in DB, Fig. 5c, d show that ESMR is superior to all the baselines in 2005 and 2007. It outperforms ECR by 3.9% and 5.5% in 2005, 24.3% and 28.9% in 2007, which obtains the best performance in all baselines. However, it is inferior to ECR by 8.8% and 13.1% in 2001 and 8.9% and 6.1% in 2003.

Similar observations can be seen by comparing ESMR with its two variations ESMR-T and ESMR-NE. The ESMR still performs better than ESMR-NE, but is worse than ESMR-T by 12.3% and 13.6% in 2001 and 4.7% and 1.2% in 2003. These are possibly due to the less sensitiveness of text topic information on paper impact in 2001 and 2003.

The results on AMiner2020 are shown in Fig. 6. The performance of all methods are better than that of AMiner2014 in different years and fields because there are more records on AMiner2020 to provide more available information and the research communities are divided by the weights provided by dataset. ESMR is still superior to the best baseline (ECR) by 14.3% and 15.5% in AI, 5.7% and 12.9% on average in DB. In addition, the overall performance is descended...
from 2001 to 2007 and from 2009 to 2015 on two AMiner datasets because it is hard to predict the influence of the new published papers. The superiority of ESMR on AMiner2014 is greater than that of AMiner2020, that shows ESMR can achieve better performance with less training data. The superiority of ESMR in 2007/2015 is more obvious than that in other years, that shows ESMR can better predict the influence of the new published papers.

### 4.6 Ranking results of author impact

Figure 7 shows the ranking results of authors on AAN dataset. We can see the performances of our ESMR and ESMR-T are better than all other methods at various k. From the overall views of average NDCG and RI, ESMR performs better than FutureRank by 41% and 42%, MRCo-Rank by 54% and 60%, ECR by 52% and 63%. Comparing
Fig. 5 The paper results of NDCG@20 and RI@20 on the AMiner2014 Dataset

ESMR with ESMR-T, ESMR still outperforms better than ESMR-T by 13.5% and 16.8% on average, which implies that adding the text information does help to better rank authors. Different from the prediction of paper impact, we have some interesting observations. LCR has achieved the worst performances except for PageRank. We guess that one possible reason is that users’ relations on paths obtained by random walk may be smaller because of a litter number of papers published by each author. Comparing with MRCoRank and FutureRank, we note that the ESMR-NE provides more competitive results, but ECR obtains relatively poor performance. The results indicate that the text topic information and the better ranking method are more important to predict author impact. From the above observations and ESMR achieving the best performances, we can conclude that it is necessary to integrate network embedding, text information and better ranking method.

Similar to the ranking of papers, we only select and rank the authors who begin to publish papers in the same research field and year on the AMiner dataset and also use NDCG and RI as metrics. In AMiner2014, the ranking results in the fields of AI and DB in 2002, 2004, 2006 and 2008 on NDCG@20 and RI@20 are listed and shown in Fig. 8a, b and Fig. 8c, d, respectively.

It can be seen that ESMR performs the best performance. For NDCG@20 and RI@20 in AI, ESMR is superior to the best baseline (ECR) by 6.4% and 9.3% on average in 2002, 2004 and 2006. In 2008, ESMR is significantly better than baselines. Comparing ESMR with ESMR-T, their results are similar in 2002 and 2004, while in 2008 ESMR is superior to ESMR-T. A possible reason is that the influence of text information added in ESMR can be propagated from papers to authors through PAN. That does help to predict the future impact of authors, especially for new authors. In the field of DB, they are 15.1% and 16.9% on average in four years. The superiority of ESMR in 2008 is not obvious, that shows it may not sensitive to text information in DB fields.

The results on AMiner2020 are shown in Fig. 9. The results are basically consistent with them on AMiner2014. The performance of all methods are also better than that of AMiner2014 because it includes more available data and the fields may be accurately divided according to the provided weights of fields in it. ESMR is still superior to ECR (the best baseline) by 14.3% and 15.5% in AI, 5.7% and 12.9% in DB on average. ESMR also significantly outperforms
other methods in 2008/2014, that proves it can better predict the impact of new authors.

### 4.7 Case study

A case study of paper prediction results on AAN dataset is presented, as shown in Table 1. In the left two columns are the index of the top-10 papers returned by the ground truth and the year of their publication. For comparison, the rankings of these papers in ESMR and the baseline approaches are listed. The boldface is used to denote that the predicting order of the papers in the top-10 list of these approaches.

Table 1 shows that ESMR gives better prediction result than all other methods. 7 papers out of top-10 papers returned by ESMR are in the top-10 papers in the ground truth, while the numbers of those hit by ECR, LCR, MRCRank, FutureRank and PageRank are 6, 5, 6, 4 and 2.

---

**Fig. 6** The paper results of NDCG@20 and RI@20 on the AMiner2020 Dataset

**Fig. 7** The author results of NDCG@$k$ and RI@$k$ on the AAN Dataset
respectively. For the influential papers P07-2045 and J07-2003 published in 2007, all the methods fail to identify them, but they ranked significantly higher in ESMR than in other methods, which indicates that ESMR improves the performance on the impact prediction of new papers.

Then a case study of ranking results of AI papers published in the year 2007 is presented. As shown in Table 2, the titles of the top-10 papers returned by ground truth and their published venues or journals are listed in the left two columns. The order of these papers in ESMR and the baselines are listed, too.

Table 2 shows that 8 out of the top-10 papers returned by ESMR are in the top-10 papers in the ground truth, while those of ECR, LCR, MRCoRank, FutureRank and PageRank are 7, 7, 5, 5 and 4, respectively. ESMR gives better ranking results than all other approaches. The influential paper *Graph Embedding and Extensions: A General Framework for Dimensionality Reduction* which ranks 2 in the ground truth is also in the top-10 list of ESMR, ECR and LCR, while the other methods fail to identify it. That is because that we only use the available papers before 2009 for ranking, its obtained citations between 2007 and 2009 is not sufficient. It demonstrates again that ESMR has powerful ability to discover the new paper with larger impact.

A case study of the author prediction results on the AAN dataset is presented. The results are shown in Table 3. The top-10 authors returned by ground truth and the future citation numbers they received are listed in the left two columns. The predicting order of these authors in ESMR and the baseline approaches are also given. Table 3 shows that 7 out of the top-10 authors returned by ESMR are in the top-10 rankings of the ground truth, while 6, 5, 3, 5 and 1 matched authors are returned by ECR, LCR, MRCoRank, FutureRank and PageRank, respectively.

Then, for AMiner2020 dataset, a case study of ranking results of authors who began to publish AI papers in 2015 are shown in Table 4. 6 out of top-10 authors returned by ESMR are in the top-10 authors in the ground truth, while the numbers of those hit by ECR, LCR, MRCoRank, FutureRank and PageRank are 5, 4, 3, 3 and 1, respectively. ESMR gives the best prediction results comparing with the baselines. For Callison-BurchChris, JurafskyDaniel in AAN and Jeff Kiske, Joel Pazhayampallil in AMiner2020, all
the methods fail to rank them, but they are significantly ranked higher in ESMR. ESMR can find new authors having scientific impact, even though they only began to publish AI papers in 2015 and have not obtained sufficient citations.

5 Discussion

In this work, we use topic information of papers, global and local structural information in HAN to predict the future

Table 1 Top 10 papers in AAN dataset

| Top-10 papers in Ground Truth | Published Year | Ranking in Different Methods |
|------------------------------|----------------|-----------------------------|
|                              |                | ESMR | ECR | LCR | MR  | FR  | PR  |
| P02-1040                     | 2002           | 6    | 7   | 3   | 4   | 4   | 14  |
| P07-2045                     | 2007           | 35   | 53  | 1384| 87  | 2379| 3949|
| P03-1021                     | 2003           | 1    | 1   | 12  | 3   | 18  | 943 |
| J03-1002                     | 2003           | 3    | 3   | 10  | 5   | 11  | 214 |
| J93-2004                     | 1993           | 4    | 4   | 6   | 1   | 7   | 118 |
| N03-1017                     | 2003           | 2    | 2   | 1   | 6   | 2   | 497 |
| J93-2003                     | 1993           | 5    | 5   | 9   | 2   | 9   | 1   |
| P03-1054                     | 2003           | 9    | 12  | 44  | 13  | 56  | 9   |
| J07-2003                     | 2007           | 62   | 98  | 166 | 140 | 238 | 5004|
| W02-1001                     | 2002           | 17   | 20  | 38  | 22  | 42  | 927 |
### Table 2  Top 10 AI papers in AMiner dataset published in 2007

| Top-10 papers in Ground Truth | Venue                  | ESMR | ECR | LCR | MR | FR | PR |
|-------------------------------|------------------------|------|-----|-----|----|----|----|
| Moses: open source toolkit for statistical machine translation | ACL        | 1    | 1   | 1   | 1  | 1  | 1  |
| Graph embedding and extensions: A general framework for dimensionality reduction | TPAMI      | 9    | 6   | 7   | 88 | 119| 37 |
| Computing semantic relatedness using Wikipedia-based explicit semantic analysis | JACAI      | 2    | 4   | 2   | 8  | 3  | 4  |
| Hierarchical phrase-based translation | Computational Linguistics | 3    | 2   | 3   | 3  | 2  | 3  |
| Local features and kernels for classification of texture and object categories: A comprehensive study | IJCV       | 5    | 10  | 9   | 4  | 5  | 13 |
| Open information extraction from the web | JACAI      | 4    | 3   | 5   | 7  | 4  | 6  |
| Tractable reasoning and efficient query answering in description logics: The DL-lite family | JAR        | 7    | 8   | 8   | 73 | 38 | 52 |
| ML-KNN: A lazy learning approach to multi-label learning | Pattern Recognition | 8    | 14  | 17  | 53 | 41 | 23 |
| MonoSLAM: Real-time single camera SLAM | TPAMI      | 70   | 101 | 157 | 39 | 478| 184|
| General tensor discriminant analysis and gabor features for gait recognition | TPAMI      | 28   | 33  | 32  | 312| 111| 546|

### Table 3  Top 10 authors in AAN dataset

| Top-10 authors in Ground Truth | future citations | ESMR | ECR | LCR | MR | FR | PR |
|-------------------------------|------------------|------|-----|-----|----|----|----|
| OchFranz Josef                | 2356             | 3    | 2   | 10  | 6  | 3  | 183|
| KleinDan                      | 2291             | 9    | 4   | 2   | 26 | 19 | 20 |
| KoehnPhilipp                  | 2119             | 5    | 11  | 20  | 3  | 18 | 143|
| ManningChristopher D.         | 1777             | 15   | 8   | 11  | 33 | 6  | 1  |
| MarcuDaniel                   | 1674             | 4    | 3   | 5   | 11 | 4  | 33 |
| KnightKevin                   | 1194             | 6    | 7   | 4   | 12 | 12 | 765|
| NeyHermann                    | 1137             | 7    | 1   | 6   | 5  | 1  | 372|
| Callison-BurchChris           | 1055             | 55   | 101 | 96  | 102| 172| 926|
| CollinsMichael John           | 1010             | 10   | 14  | 26  | 20 | 5  | 796|
| JurafskyDaniel                | 955              | 44   | 21  | 35  | 135| 25 | 1079|

### Table 4  Top 10 authors began to publish AI papers in AMiner2020 dataset

| Top-10 authors in Ground Truth | future citations | ESMR | ECR | LCR | MR | FR | PR |
|-------------------------------|------------------|------|-----|-----|----|----|----|
| Daniel Horgan                 | 31               | 4    | 3   | 8   | 7  | 5  | 198|
| Anh Nguyen                    | 27               | 3    | 5   | 4   | 35 | 16 | 29 |
| Thomas J. fuchs               | 27               | 2    | 13  | 24  | 9  | 41 | 167|
| Bradly C. Stadie              | 22               | 13   | 9   | 6   | 42 | 56 | 9  |
| Brody Huval                   | 18               | 15   | 23  | 31  | 11 | 8  | 33 |
| Tao Wang                      | 18               | 9    | 6   | 16  | 28 | 19 | 378|
| Sameep Tandon                 | 18               | 5    | 2   | 6   | 4  | 15 | 269|
| Jeff Kiske                    | 18               | 43   | 96  | 114 | 178| 209| 762|
| Will Song                     | 18               | 8    | 17  | 31  | 23 | 6  | 826|
| Joel Pazhayampallil           | 17               | 31   | 42  | 38  | 173| 36 | 968|
scientific impact. We can construct a more comprehensive HAN in the future works, which includes much other information, such as publication venue, journal, publisher, the institutions of authors and so on, and extend ESMR for getting better ranking results.

For Aminer2014, the papers are divided into different research fields in term of published venues. This may be unreasonable because that papers with different topics can be published in the same published venue or papers with same topic can be published in different published venues. We will divide them by extracting keywords of abstract and title content, which may refine the field division of papers.

In our experiments, the papers are divided into the training and test set based on whether or not they published before 2009/2015. Since this may affect the prediction results, in the follow-up research, we will consider to divide the data with other time points, and train the corresponding models. In addition, we can replace two-stage training with collaborative training to explore further better predicting results.

On the other hand, although the proposed ESMR cannot completely solve the problem of predicting the future scientific impact, it is the first attempt to learn a unified entity representations by using network embedding based model, and to integrate richer information to a multivariate random-walk model for improving the prediction performance.

6 Conclusions

In this paper, a new ranking method ESMR was proposed to predict the future scientific influence of papers and authors. A network embedding based model was designed to learn a unified better representations for various entities in the constructed heterogeneous academic network. The learned embedding representations could capture the local structural information, the rich text and time information of papers, which is important to effectively predict the scientific impact. By integrating the learned embeddings and the global structural information into a multivariate random-walk model, the future impact of papers and authors were predicted simultaneously, especially for new papers and authors. The experimental results on two datasets demonstrated that the proposed model outperformed other baselines.

Abbreviations  HAN, Heterogeneous Academic Network; ESMR, Embeddings and Structure information to Multivariate Random-walk; PCN, Paper Citation Network; PCNT, Paper Citation Network with Text information; TINp, Text Information Network of a Paper; PAN, Paper Author Network; CAN, Co-Author Network; AAN, ACL Anthology Network; RI, Recommendation Intensity; NDCG, Normalized Discounted Cumulative Gain; AI, Artificial Intelligence; DB, DataBase.

Acknowledgements  This work was supported by the National Natural Science Foundation of China (U1533104; U1933114) and National Key Laboratory Open Foundation of Air Traffic Management System and Technology (SKLATM201902).

Author Contributions  CX participated in model design and helped to draft the manuscript. LS participated in the study of related work and helped to draft the manuscript. JH carried out the training and testing of the method and helped to draft the manuscript. CX participated in model design and helped to draft the manuscript. All authors read and approved the final manuscript. YQ participated in the construction of Academic Network and helped to draft the manuscript.

Funding  The National Natural Science Foundation of China (U1533104; U1933114) and National Key Laboratory Open Foundation of Air Traffic Management System and Technology (SKLATM201902).

Data Availability  The datasets used and analysed during the current study are available from https://github.com/PercyShelly/ESMR.git.

Declarations

Competing interests  The authors declare that they have no competing interests.

References

1. Zhang L, Fan Y, Zhang W, Zhang S, Yu D, Zhang S (2019) Measuring scientific prestige of papers with time-aware mutual reinforcement ranking model. Journal of Intelligent and Fuzzy Systems 36(2):1–15
2. Fiša J, Sesták M, Jiri J (2017) Reflections on how to evaluate the professional value of scientific papers and their corresponding citations. Scientometrics: An International Journal for All Quantitative Aspects of the Science of Science Policy
3. Wang Y, Yunhai T, Zeng M (2013) Ranking scientific articles by exploiting citations, authors, journals, and time information. In: Proceedings of AAAI. AAAI, pp 933–939
4. Khan S, Liu X, Shakkil KA, Alam M (2017) A survey on scholarly data: from big data perspective. Inf Process Manag 53(4):923–944
5. Liang R, Jiang X (2016) Scientific ranking over heterogeneous academic hypernetwork. In: Proceedings of AAAI, pp 20–26
6. Xia F, Wang W, Bekele TM, Liu H (2017) Big scholarly data: a survey. IEEE Transactions on Big Data 3(1):18–35
7. Hirsch J. E (2019) hα: An index to quantify an individual’s scientific leadership. Scientometrics 118(2):673–86
8. Silagadze ZK (2009) Citation entropy and research impact estimation. Acta Phys Polon 4(11):2325–2333
9. Xu J, Mengjun L, Jiang J, Ge B (2019) Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network. IEEE Access PP(99)
10. Jiang X, Gao C, Liang R (2017) Ranking scientific articles in a dynamically evolving citation network. In: International conference on semantics
11. Sayyadi H, Getoor L (2009) Futurerank: ranking scientific articles by predicting their future pagerank. In: Proceedings of SDM, pp 533–544
12. Wang S, Xie S, Zhang X, Li Z, He Y, He Y (2016) Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement. ACM Trans Intell Syst Technol 7(4):64
13. Wang S, Xie S, Zhang X, Li Z, Yu PS, Shu X (2014) Future influence ranking of scientific literature. In: Proceedings of SDM, pp 749–757

14. Zhang J, Xia F, Wang W, Bai X, Yu S, Bekele TM, Peng Z (2016) Cocarank: a collaboration caliber-based method for finding academic rising stars. In: Proceedings of WWW, pp 395–400

15. Xiao C, Han J, Fan W, Wang S, Huang R, Zhang Y (2019) Predicting scientific impact via heterogeneous academic network embedding. In: Pacific Rim international conference on artificial intelligence. Springer, pp 555–568

16. Wu Z, Lin W, Liu P, Chen J, Mao L (2019) Predicting long-term scientific impact based on multi-field feature extraction. IEEE Access 7:51759–51770

17. Abramo G, D’Angelo CA, Felici G (2019) Predicting publication long-term impact through a combination of early citations and journal impact factor. Journal of Informetrics 13(1):32–49

18. Chaturvedi S (2016) Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology 67(11):2684–2696

19. Dong Y, Johnson RA, Chawla NV (2016) Can scientific impact be predicted? IEEE Transactions on Big Data 2(1):18–30

20. Hayashi T, Fujita H (2021) Cluster-based zero-shot learning for multivariate data. Journal of Ambient Intelligence and Humanized Computing 12:1897–1911

21. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q (2015) LINE: large-scale information network embedding. In: Proceedings of WWW, pp 1067–1077

22. Xu L, Wei X, Cao J, Yu PS (2017) Embedding of embedding (EOE): joint embedding for coupled heterogeneous networks. In: Proceedings of WSDM, pp 741–749

23. He Z, Liu J, Zeng Y, Wei L, Huang Y (2021) Content to node: Self-translation network embedding. IEEE Transactions on Knowledge and Data Engineering 33(2):431–443

24. Bras-Amorós M, Domingo-Ferrer J, Torra V (2011) A bibliometric index based on the collaboration distance between cited and citing authors. Journal of Informetrics 5(2):248–264

25. Zhou D, Orshanskiy SA, Zha H, Giles CL (2007) Co-ranking authors and documents in a heterogeneous network. In: Proceedings of ICDM, pp 739–744

26. Wang Z, Chen C, Li W (2017) Predictive network representation learning for link prediction. In: Proceedings of SIGIR, pp 969–972

27. Hayashi T, Fujita H, Hernandez-Matamoros A (2021) Less complexity one-class classification approach using construction error of convolutional image transformation network. Information Sciences 560:217–234

28. Cavallari S, Zheng VW, Cai H, Chang KC-C, Cambria E (2017) Learning community embedding with community detection and node embedding on graphs. In: Proceedings of CIKM, pp 377–386

29. Hu B, Shi C, Zhao WX, Yu PS (2018) Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In: Proceedings of KDD, pp 1531–1540

30. Campbell JC, Hindle A, Stroufia E (2015) Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data. The Art and Science of Analyzing Software Data. 2015:139–159

31. Qin P, Xu W, Guo J (2016) A novel negative sampling based on tfidf for learning word representation. Neurocomputing 177:257–265

32. Liu X (2018) A weighted pagerank for scientific paper ranking. In: Proceedings of 2018 the 8th international workshop on computer science and engineering, pp 136–140

Publisher's note Springer Nature remains neutral with regard to jurisdicational claims in published maps and institutional affiliations.