RESEARCH

Heterogeneous Academic Network Embedding Based Multivariate Random-Walk Model for Predicting Scientific Impact

Yongwei QIAO[1*], Leilei SUN[2], Jianing HAN[3] and Chunjing XIAO[2]

[*]Correspondence:
qiaoyongwei76@126.com
[1]Engineering Technology Training Center, Civil Aviation University of China, Jinbei Road, 300300 Tianjin, china
Full list of author information is available at the end of the article

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

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, and direct policymakers in recruitment decisions or funding allocation [1–3]. So far, most remarkable works have focused on ranking the *current importance* of papers and authors [2, 4], and proposed some more complicated metrics, such as h-index [5] and s-index [6].

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 [5, 7]. 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

Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

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 [8], 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, 9–12].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 [2, 10]. 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, 4]. 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. The *local structure information* (i.e., local similarity between entities) and the *text information of papers* (i.e., paper topics), which is helpful to identify influential papers and authors, can not be directly taken into account [13]. 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 reason is that their structure information of citation and co-author are sparse and not insufficient to characterize their innovativeness). 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 [10, 14], author order factors [14], early citations [15], journal impact factor [15], text features [10, 16], topical authority [15, 17], and so on. But it is difficult to express the information well in a unified representation to integrate those into prediction approaches. Recent advances based on network embedding [18–20] 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 **E**mbeddings and global **S**tructure information to **M**ultivariate **R**andom-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 topic information simultaneously. Then the future scientific impact of papers and authors is comprehensively predicted by integrating the learned embeddings and global structural information into the multivariate random walk

algorithm. Extensive experiments on two datasets are 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.

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 [5], g-index [7], c-index [21] and s-index [6]. 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 [8]. Although they can give the current influence of papers or authors, it is difficult to predict the future influence of them.

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 [2, 13]. The *graph-based models* including univariate and multivariate random walk techniques are widely used to predict future impact of papers [3, 9–12]. 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) [2, 10]. Such straightforward methods ignore the different influences between different types of objects, and thus limit the effectiveness in ranking different kinds of objects. 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 [22] 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) [2, 4, 9, 10].

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. [9] and Wang et al. [10] 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. [14] proposed TAORank, which considers mutual influence among scholarly entities, which includes the publication time and author order in scholarly papers. Wang et al. [10] incorporated the text information because they thought that it is useful to improve the predicting results. Chaturvedi and Snigdha [16] 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. [17] and Giovanni et al. [15] found that topical authority and publication venue were crucial to these effective predictions. Liang et al. [4] 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 [23], node classification [18, 20], community detection [24], and recommendation task [25]. 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.

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 Figure 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 is preserved; (3) predicting the future impact of papers and authors by integrating these learned representations into a multivariate random walk based Co-Ranking model.

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.*

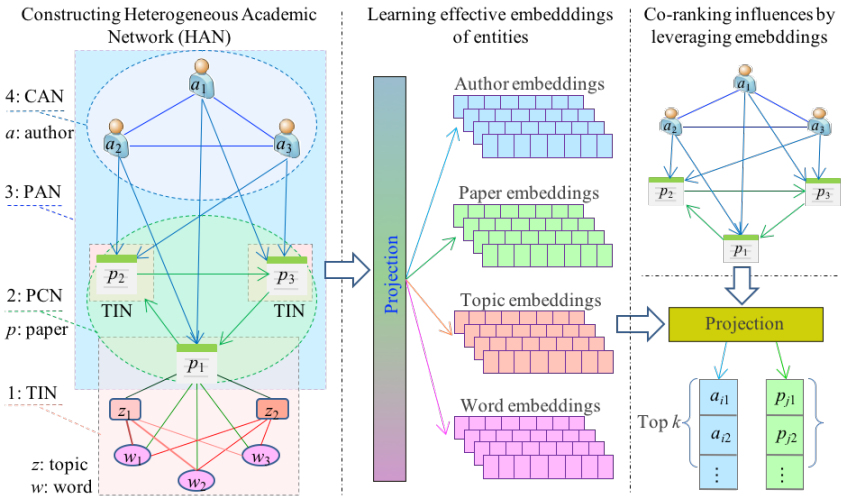


Figure 1 The framework of ESMR. The process of ESMR consists of three parts: (1) constructing the HAN; (2) learning unified Embedding representations of entities in HAN; (3) predicting the future influence of papers and authors.

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 $p_i \in P$, its text information network is denoted as a three-tuple $G_p^i = (V_p^i, E_p^i, F_p^i)$, where V_p^i denotes the nodes including the words and topics contained in paper p_i, E_p^i denotes edges, F_p^i is edge weights, respectively.*

TINp of a paper p_i contains paper-topic network G_{pz}^i, paper-word network G_{pw}^i and topic-word network G_{zw}^i. Note that $V_p^i = \{p_i \cup Z_p^i \cup W_p^i\}$, where Z_p^i and W_p^i are the topics and words of p_i, respectively; $E_p^i = \{E_{pz}^i \cup E_{pw}^i \cup E_{zw}^i\}$, where E_{pz}^i and E_{pw}^i are the edges between p_i and its topics and words, E_{pw}^i is the edges between topics and words of p_i, respectively; $F_p^i = \{F_{pz}^i \cup F_{pw}^i \cup F_{zw}^i\}$, where F_{pz}^i, F_{pw}^i and F_{zw}^i are the weights of edges E_{pz}^i, E_{pw}^i and E_{zw}^i, respectively. 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) [13].*** *Let 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, respectively. The PCNT is a combination of the different types of vertices and relations, which is defined as $G_p = (V_p, E_p, F_p)$, where $V_p = \{P \cup Z \cup W\}$, $E_p = \{E_{pp} \cup E_{pz} \cup E_{pw} \cup E_{zw}\}$, and $F_p = \{F_{pp} \cup F_{pz} \cup F_{pw} \cup F_{zw}\}$. F_{pp}, F_{pz}, F_{pw} and 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.*

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}.*

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) [13].*** *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 edges.*

Embedding for Heterogeneous Academic Network

Embedding for PCNT.

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

$$\mathcal{L}_p = -\sum_{(i,j)\in E_p} \lambda_p^i \hat{\mathcal{P}}(v_j|p_i) \log \mathcal{P}(v_j|p_i), \tag{1}$$

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

$$\mathcal{P}(v_j|p_i) = \frac{\exp({\mathbf{v}_j}^\top \cdot \mathbf{p}_i)}{\sum_{k\in V_p} \exp({\mathbf{v}_k}^\top \cdot \mathbf{p}_i)}, \tag{2}$$

where $\mathbf{p}_i \in \mathcal{R}^d$ and $\mathbf{v}_j \in \mathcal{R}^d$ are d-dimensional latent representations of p_i and v_j, respectively.

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

For G_{pp}, $v_j \in P$, $R(p_i)$ is the set of papers cited by p_i, and ω_{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 ω_{ij} is defined as:

$$\omega_{ij} = e^{-\rho(T_c - T_{i\rightarrow j})}, \tag{3}$$

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

For G_{pz}, $v_j \in Z$, $R(p_i)$ is the set of topics that are most likely to be touched upon p_i. ω_{ij} ($\omega_{ij} \in F_{pz}$) 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 [26]. And λ_p^i is defined as $\lambda_{pz}^i = \sum_{k\in R(p_i)} \omega_{ik}$. 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. ω_{ij} ($\omega_{ij} \in F_{pw}$) reflects the importance of w_j in p_i and can be calculated by if-idf. And λ_p^i is denoted as $\lambda_{pw}^i = \sum_{k\in R(p_i)} \omega_{ik}$. It represents the influence of paper p_i over the words.

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:

$$\mathcal{L}_a = -\sum_{(i,j)\in E_a} \lambda_a^i \hat{\mathcal{P}}(a_j|a_i) \log \mathcal{P}(a_j|a_i). \tag{4}$$

where $\hat{\mathcal{P}}(a_j|a_i) = \frac{\omega_{ij}^a}{\sum_{k\in N(a_i)} \omega_{ik}^a}$, $N(a_i)$ is the set of co-authors of author a_i, and ω_{ik}^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

$$\omega_{ij}^{a} = \sum_{p_k \in Co(a_i,a_j)} e^{-\rho(T_c - T_{co}^{p_k})}, \tag{5}$$

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

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 ω_{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 $\mathcal{P}(p_i, a_j)$ 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.

$$\mathcal{L}_{pa} = -\sum_{(i,j) \in E_{pa}} \omega_{ij}^{pa} \log \mathcal{P}(p_i, a_j), \tag{6}$$

where ω_{ij}^{pa} is set to $\frac{1}{s}$, and s is the signature order of author a_i in paper p_j. $\mathcal{P}(p_i, a_j)$ is defined as $\mathcal{P}(p_i, a_j) = \frac{1}{1+\exp(-\mathbf{p}_i^\top \cdot \mathbf{a}_j)}$, where $\mathbf{p}_i \in \mathcal{R}^d$ and $\mathbf{a}_j \in \mathcal{R}^d$ are d-dimensional latent representations of p_i and a_j, respectively.

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 Eq. (1), Eq. (4) with Eq. (6), then jointly minimize the following objective function.

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_a + \mathcal{L}_{pa}. \tag{7}$$

Model Optimization.
Direct calculation of the function in Eq.(2) [18] is both time-consuming and impractical. To address the computational challenge, negative sampling approach [27] 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 $\mathcal{L}$ can be expressed as the following formula, which uses L2-norm to avoid over-fitting and ignores some constraints.

$$\begin{aligned} \mathcal{L} = &-\sum_{(i,j) \in E'} \lambda_v^i \hat{\mathcal{P}}(v_j|v_i) \log(\sigma({\mathbf{v}_j}^\top \cdot \mathbf{v}_i)) - \sum_{(i,j) \notin E'} \lambda_v^i \log(\sigma(-{\mathbf{v}_j}^\top \cdot \mathbf{v}_i)) \\ &- \sum_{(i,j) \in E_{pa}} \omega_{pa}^{ij} \log(\sigma(\mathbf{p}_i^\top \cdot \mathbf{a}_j)) + \lambda \sum_{n=1}^{|P|} ||\mathbf{p}_n||_2 + \beta \sum_{n=1}^{|V_a|} ||\mathbf{a}_n||_2, \end{aligned} \tag{8}$$

where $\sigma(x) = \frac{1}{1+exp(-x)}$ is the sigmoid function. λ, $\beta \in \mathcal{R}$ are regularization coefficients. The positive samples are modeled by the first term in Eq. (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 Eq.(7), the gradients for $\mathbf{p}_i$ and $\mathbf{a}_i$ can be computed by using the stochastic gradient descent algorithm.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{p}_i} =& -\sum_{(i,j)\in E_p} \lambda_p^i \frac{\hat{\mathcal{P}}(v_j|p_i) exp\{-{\mathbf{v}_j}^\top \mathbf{p}_i\}}{1+exp\{-{\mathbf{v}_j}^\top \mathbf{p}_i\}} \mathbf{v}_j + \sum_{(i,j)\notin E_p} \frac{\lambda_p^i}{1+exp\{-{\mathbf{v}_j}^\top \mathbf{p}_i\}} \mathbf{v}_j \\ & - \sum_{(i,j)\in E_{pa}} \omega_{pa}^{ij} \frac{exp\{-\mathbf{p}_i^\top \mathbf{a}_j\}}{1+exp\{-\mathbf{p}_i^\top \mathbf{a}_j\}} \mathbf{a}_j + \lambda \sum_{r=1}^{d_p} 2(\mathbf{p}_i^r),\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{a}_i} =& -\sum_{(i,j)\in E_a} \lambda_a^i \frac{\hat{\mathcal{P}}(a_j|a_i) exp\{-{\mathbf{a}_j}^\top \mathbf{a}_i\}}{1+exp\{-{\mathbf{a}_j}^\top \mathbf{a}_i\}} \mathbf{a}_j + \sum_{(i,j)\notin E_a} \frac{\lambda_a^i}{1+exp\{-{\mathbf{a}_j}^\top \mathbf{a}_i\}} \mathbf{a}_j \\ & - \sum_{(j,i)\in E_{pa}} \omega_{pa}^{ji} \frac{exp\{-\mathbf{p}_j^\top \mathbf{a}_i\}}{1+exp\{-\mathbf{p}_j^\top \mathbf{a}_i\}} \mathbf{p}_j + \beta \sum_{r=1}^{d_a} 2(\mathbf{a}_i^r),\end{aligned} \tag{10}$$

where d_p and d_a respectively represent the dimension of vectors $\mathbf{p}$ and $\mathbf{a}$, and here $d_p = d_a = d$. We do not detailedly show the gradients for $\mathbf{z}_i$ and $\mathbf{w}_i$, which can be derived in the similar way.

For each iteration, we adopt the backtracking line search [19] 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 $O(nd_p \times d_a k)$. Therefore, it is easy to see that the training can be done in polynomial time. The detailed training process has shown in Algorithm 1 [13].

Algorithm 1 The training process for embedding HAN.

Input: $G(V,E,F)$, learning rate η, dimensions of vectors d, negative sampling rate k, regularized coefficients λ and β
Output: latent vectors of papers, topics, words and authors
initializing vectors of $\mathbf{p}, \mathbf{z}, \mathbf{w}$ and $\mathbf{a}$;
while (not converged) **do**
    for all P, calculating $\frac{\partial \mathcal{L}_p}{\partial \mathbf{p}_i}$;
    setting up η_i by using backtracking line search;
    for all P, updating $\mathbf{p}$;
    for all P and Z, updating $\mathbf{p}$ and $\mathbf{z}$;
    for all P and W, updating $\mathbf{p}$ and $\mathbf{w}$;
end while
for all V_a, pre-training $\mathbf{a}$;
while (not converged) **do**
    optimizing $V_p(\mathbf{p}, \mathbf{z}, \mathbf{w})$ based on Eq. (9) and $\frac{\partial \mathcal{L}}{\partial \mathbf{z}_i}$, $\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i}$ when fix V_a;
    similarly optimizing$V_a(\mathbf{a})$ based on Eq. (10) when fix V_p;
end while
return latent vectors of $\mathbf{p}, \mathbf{z}, \mathbf{w}$ and $\mathbf{a}$

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 $Sim(p_i, p_j) = \frac{\mathbf{p}_i \cdot \mathbf{p}_j}{||\mathbf{p}_i|| \times ||\mathbf{p}_j||}$. Thus, to be used in the multivariate random-walk model, the transition matrix of PCN can be represented as:

$$M_{pp}(j,i) = \begin{cases} \gamma \frac{Sim(p_i,p_j)}{\sum_{p_k \in N(p_i)} Sim(p_i,p_k)} + (1-\gamma)\frac{\lambda^i_{pp}}{deg(p_i)}, & e^{ij}_{pp} \in E_{pp} \\ 0, & otherwise \end{cases} \tag{11}$$

where $N(p_i)$ and $deg(p_i)$ respectively represent the sets of neighbors and out-degree nodes of p_i, γ 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 calcuate the future influence of papers and authors. Each iteration process is defined as the following equations:

$$\mathbf{p}^{(t+1)} = \alpha_{pp} M_{pp} \mathbf{p}^{(t)} + \beta_{pa} M_{pa} \mathbf{a}^{(t)}, \tag{12}$$

$$\mathbf{a}^{(t+1)} = \alpha_{aa} M_{aa} \mathbf{a}^{(t)} + \beta_{ap} M_{ap} \mathbf{p}^{(t)}, \tag{13}$$

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

Experimental Results.

Datasets.

ESMR is evaluated by using the following two public datasets. One is the ACL Anthology Network (AAN) [28], 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[1] [29]. It is one of the datasets released by AMiner. The dataset includes 2,092,356 papers published before 2014 and their 8,024,869 citations. The metadata for each paper contains the following information: paper ID, paper title, author list, author affiliation, published year, published venue, abstract and the list of references.

Firstly, the dataset is preprocessed as follows. Papers published after 1998 were selected for evaluating the predicting performance. Then the papers without sufficient metedata, 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. While in the AMiner dataset, we obtain 328,971 papers and 2,732,340 citations.

[1]https://www.aminer.cn/aminernetwork

Experiment Setup.

Ground Truth.

Owing to the lack of criteria, it is a challenge to evaluate the performance of almost all works. Following recent works [3, 10], the number of future citations is used as the ground truth to evaluate it.

The dataset is divided into two parts according to a historical time point. The training part is used to obtain the estimated ranking lists of papers and authors. 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 according to the future citation number. Finally, the results is reported by comparing the similarities of the two rank lists.

In this paper, the papers are divided into the training and test set based on whether or not they published before 2009. The future citations of papers published before 2009 are calculated in periods 2010, 2010-2012, and 2010-2014. For example, the results in the AMiner dataset 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 2010-2014 are regarded as the ground truth to evaluate our prediction results.

Evaluation Metrics.

There are two widely used metrics to evaluate the performance. One is recommendation intensity RI [2, 10, 30]. 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_r, the recommendation intensity of P_i at k can be defined as

$$RI(P_i)@k = \begin{cases} 1 + (k - o_r)/k, & P_i \in L \\ 0, & P_i \notin L \end{cases} \tag{14}$$

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 [10, 30], 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.

$$NDCG@k = \frac{DCG@k}{IDCG@k} = \frac{\sum_{i=1}^{k} \frac{2^{rel_i}-1}{\log_2(i+1)}}{\sum_{i=1}^{|REL|} \frac{2^{rel_i}-1}{\log_2(i+1)}}. \quad (15)$$

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.

Baselines.

To evaluate ESMR, the following methods are used to compare with it.

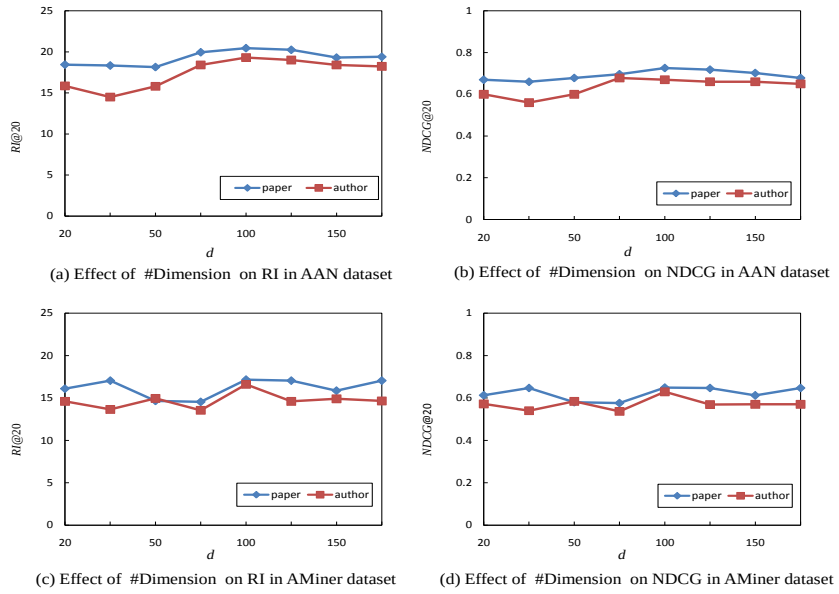
- **MRCoRank (MR).** MRCoRank is the state-of-the-art graph-based method to rank the future influence of papers, authors, and venues simultaneously. It takes time, text and structure information into account when using a mutual reinforcement framework to predict the results [10].
- **FutureRank (FR).** 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) [9].
- **PageRank (PR).** PageRank is a base model for many graph-based ranking methods [31], 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 [18]. 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 [19] 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.

Parameter Sensitivity Analysis.

In the process of training ESMR, the negative sampling rate is set to 5. The regularization coefficients λ, β 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 Figure 2 (a)-(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. So, 100 is used as the default dimension setting in the following experiments.

Furthermore, here are four parameters in the process of ranking, α_{pp}, α_{aa}, β_{pa} and β_{ap}. Take the AAN dataset as an example, Figure 3 (a)-(b) shows the effect of α_{pp} on papers, and Figure 3 (c)-(d) shows the effect of α_{aa} on authors. For α_{pp}, it is set to 0.3 for getting the best result. For α_{aa}, 0.15 is a reasonable choice. Whether it is too large or too small for both α_{pp} and α_{aa} will reduce the performance. In the following experiments, for the AAN dataset, $\alpha_{pp} = 0.3$, $\alpha_{aa} = 0.15$, $\beta_{pa} = 0.3$



(a) Effect of #Dimension on RI in AAN dataset

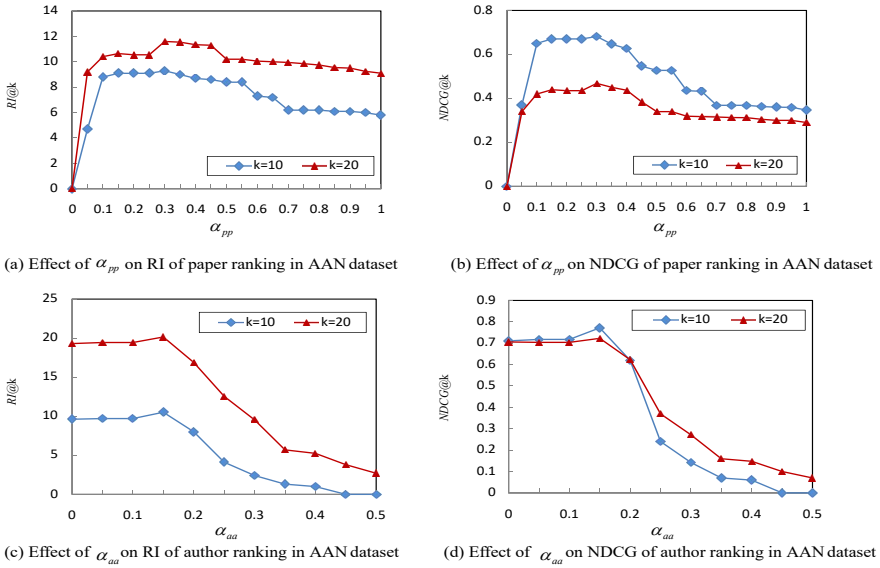
(b) Effect of #Dimension on NDCG in AAN dataset

(c) Effect of #Dimension on RI in AMiner dataset

(d) Effect of #Dimension on NDCG in AMiner dataset

Figure 2 Analysis of dimensions of vectors.

and $\beta_{ap} = 0.85$ are the default parameter settings. And for the AMiner dataset, 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 γ in Eq.(11) while for the AMiner dataset, the value is 0.6.



(a) Effect of α_{pp} on RI of paper ranking in AAN dataset

(b) Effect of α_{pp} on NDCG of paper ranking in AAN dataset

(c) Effect of α_{aa} on RI of author ranking in AAN dataset

(d) Effect of α_{aa} on NDCG of author ranking in AAN dataset

Figure 3 Sensitivity analysis for α_{pp} and α_{aa}.

The parameters of the 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

$\gamma v = \gamma_{vp} = \gamma_{va} = 0.4$. 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.

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 Figure 4.

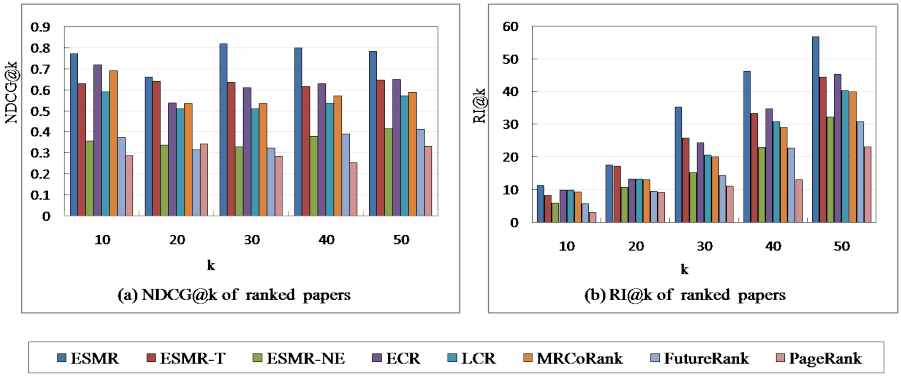


Figure 4 The paper results of NDCG@k and RI@k on the AAN Dataset.

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 the two metrics by 33% and 42% over MRCoRank which achieves the best performance in all baselines. 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 of paper impact prediction. ESMR-T is consistent with ECR in most cases, sometimes better than ECR. The possible reason is that they all 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 number of papers published before 2006 account for more than 80%. That is to say, the ground truth obviously tends to the older ones. (2) The difference in the number of future citations is very large for the papers published in various kinds of research fields. For example, the future citations of the most cited papers in the field of Information Security is

245, while there are 37 papers in Artificial Intelligence obtaining more than 245 citations. Therefore, in order to give 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 2001, 2003, 2005 and 2007 are selected to evaluate the paper ranking performance. It is shown in Figure 5 (a)-(b) and NDCG@20 and RI@20 are taken as metrics, respectively.

The results on Aminer dataset are basically consistent with the results on AAN, but there are still some differences. For the ranking results in AI, 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 MRCoRank by 31% and 40% in 2003, than LCR by 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, Figure 5 (c)-(d) show that ESMR is superior to all the baselines in 2005 and 2007, but is inferior to ECR in 2001 and 2003.

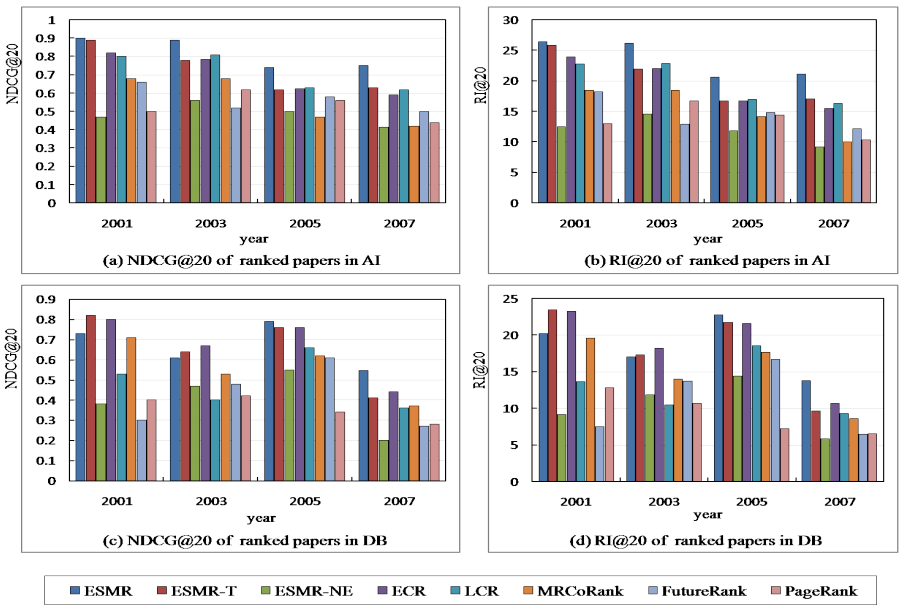


Figure 5 The paper results of NDCG@20 and RI@20 on the AMiner Dataset.

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 in 2001 and 2003. These are possibly due to the less sensitiveness of text topic information on paper impact in 2001 and 2003.

Ranking Results of Author Impact.

Figure 6 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

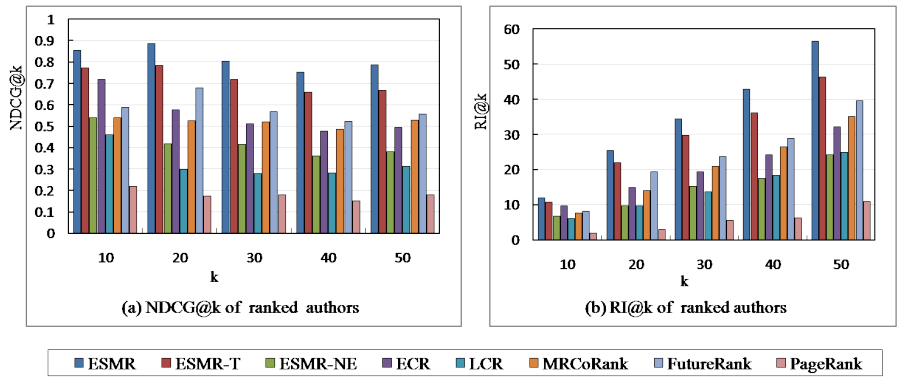


Figure 6 The author results of NDCG@k and RI@k on the AAN Dataset

than FutureRank by 41% and 42%, MRCoRank by 54% and 60%, ECR by 52% and 63%. Comparing ESMR with ESMR-T, ESMR still outperforms better than ESMR-T, 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 the 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. For NDCG@20 and RI@20 in the fields of AI and DB, the ranking results in 2002, 2004, 2006 and 2008 are listed and shown in Figure 7 (a)-(b) and Figure 7 (c)-(d), respectively. It can be seen that ESMR performs the best performance. For NDCG@20 and RI@20 in AI, ESMR is superior to ECR by 11% and 15% in 2006. In 2008, ESMR is significantly better than baselines. In the feld of DB, they are 37%, 30% and 8%, 11% in the year of 2006 and 2008, respectively. 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.

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 figure is used to denote that the predicting order of the papers in the top-10 list of these approaches.

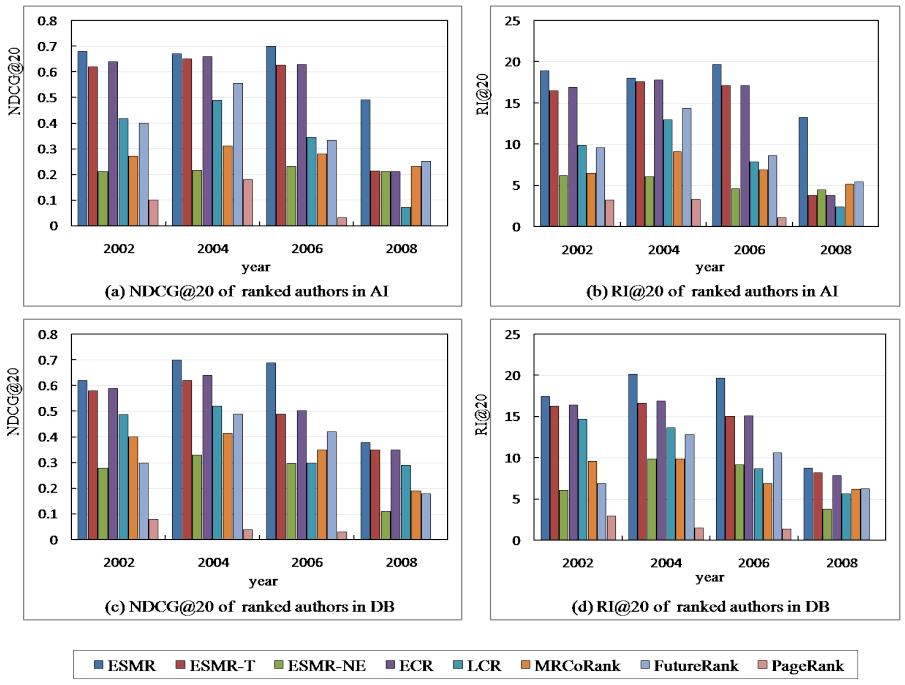


Figure 7 The author results of NDCG@20 and RI@20 on the AMiner Dataset

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 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, MRCoRank, FutureRank and PageRank are 6, 5, 6, 4 and 2, 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. And 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. The influential paper *Graph Embedding*

and Extensions: A General Framework for Dimensionality Reduction which ranks 2 in the groud 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. ESMR gives better ranking results than all the other approaches. It demonstrates again that ESMR has powerful ability to discover the new paper with lager impact.

Table 2 Top 10 AI papers in AMiner dataset published in 2007.

Top-10 papers in Ground Truth	Venue	Ranking in Different Methods					
		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	IJCAI	**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	IJCAI	**4**	**3**	**5**	**7**	**4**	**6**
Tractable reasoning and efficient query answering in description logics: The DL-lite family	J Autom Reasoning	**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

A case study of the author prediction results on the AAN dataset is also 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

Table 3 Top 10 authors in AAN dataset

Top-10 authors in Ground Truth	future citations	Ranking in Different Methods					
		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

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.

ESMR gives the best prediction result comparing with the baselines. Similar to the ranking results of papers on the AMiner dataset, ESMR can find new authors having scientific impact, even though they only began to publish papers in 2006 and have not obtained sufficient citations.

Discussion

In this work, we use topic information of papers, global and local structural information in HAN to predict the future scientific impact. We can construct a more comprehensive HAN in the future work, 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 Aminer dataset, 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 experiment, the papers are divided into the training and test set based on whether or not they published before 2009. 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.

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.

Declarations

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

Availability of data and materials
The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Funding
The National Natural Science Foundation of China (U1533104;U1933114)

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

Author's contributions
YQ conceived of the study, participated in the construction of Academic Network 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.

Acknowledgements
This work was supported by grants from the National Natural Science Foundation of China (U1533104; U1933114)

Author details
[1]Engineering Technology Training Center, Civil Aviation University of China, Jinbei Road, 300300 Tianjin, china. [2]College of Computer Science and Technology, Civil Aviation University of China, Jinbei Road, 300300 Tianjin, China. [3]Information Management Department, Air China Stock Corporation, Tianzhu West Road, 101300 Beijing, china.

References

1. Garfield, E.: Citation analysis as a tool in journal evaluation. Science **178**(4060), 471–479 (1972)
2. Jiang, X., Sun, X., Zhuge, H.: Towards an effective and unbiased ranking of scientific literature through mutual reinforcement. In: Proceedings of CIKM, pp. 714–723 (2012)
3. Wang, Y., Yunhai, T., Zeng, M.: Ranking scientific articles by exploiting citations, authors, journals, and time information. In: Proceedings of AAAI, pp. 933–939 (2013)
4. Liang, R., Jiang, X.: Scientific ranking over heterogeneous academic hypernetwork. In: Proceedings of AAAI, pp. 20–26 (2016)
5. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences **102**(46), 16569–16572 (2005)
6. Silagadze, Z.K.: Citation entropy and research impact estimation. Acta Physica Polonica **41**(11) (2009)
7. Egghe, L.: Theory and practice of the g-index. Scientometrics **69**(1), 131–152 (2006)
8. Jiang, X., Gao, C., Liang, R.: Ranking Scientific Articles in a Dynamically Evolving Citation Network. In: 154-157 (ed.) International Conference on Semantics, Knowledge and Grids (2017)
9. Sayyadi, H., Getoor, L.: Futurerank: Ranking scientific articles by predicting their future pagerank. In: Proceedings of SDM, pp. 533–544 (2009)
10. Wang, S., Xie, S., Zhang, X., Li, Z., He, Y., He, Y.: Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement. Acm Trans. on Intelligent Systems & Technology **7**(4), 64 (2016)
11. Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P.S., Shu, X.: Future influence ranking of scientific literature. In: Proceedings of SDM, pp. 749–757 (2014)
12. Zhang, J., Xia, F., Wang, W., Bai, X., Yu, S., Bekele, T.M., Peng, Z.: Cocarank: A collaboration caliber-based method for finding academic rising stars. In: Proceedings of WWW, pp. 395–400 (2016)
13. Xiao, C., Han, J., Fan, W., Wang, S., Huang, R., Zhang, Y.: Predicting scientific impact via heterogeneous academic network embedding. In: Pacific Rim International Conference on Artificial Intelligence, pp. 555–568 (2019). Springer
14. Wu, Z., Lin, W., Liu, P., Chen, J., Mao, L.: Predicting long-term scientific impact based on multi-field feature extraction. IEEE Access **7**, 51759–51770 (2019)
15. Abramo, G., D'Angelo, C.A., Felici, G.: Predicting publication long-term impact through a combination of early citations and journal impact factor. Journal of Informetrics **13**(1), 32–49 (2019)
16. Chaturvedi, S.: Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology **67**(11), 2684–2696 (2016)
17. Dong, Y., Johnson, R.A., Chawla, N.V.: Can scientific impact be predicted? IEEE Transactions on Big Data **2**(1), 18–30 (2016)
18. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: LINE:large-scale information network embedding. In: Proceedings of WWW, pp. 1067–1077 (2015)
19. Xu, L., Wei, X., Cao, J., Yu, P.S.: Embedding of Embedding (EOE): Joint embedding for coupled heterogeneous networks. In: Proceedings of WSDM, pp. 741–749 (2017)
20. Liu, J., He, Z., Wei, L., Huang, Y.: Content to node: Self-translation network embedding. In: Proceedings of KDD, pp. 1794–1802 (2018)
21. Bras-Amorós, M., Domingo-Ferrer, J., Torra, V.: A bibliometric index based on the collaboration distance between cited and citing authors. Journal of Informetrics **5**(2), 248–264 (2011)
22. Zhou, D., Orshanskiy, S.A., Zha, H., Giles, C.L.: Co-ranking authors and documents in a heterogeneous network. In: Proceedings of ICDM, pp. 739–744 (2007)

23. Wang, Z., Chen, C., Li, W.: Predictive network representation learning for link prediction. In: Proceedings of SIGIR, pp. 969–972 (2017)
24. Cavallari, S., Zheng, V.W., Cai, H., Chang, K.C.-C., Cambria, E.: Learning community embedding with community detection and node embedding on graphs. In: Proceedings of CIKM, pp. 377–386 (2017)
25. Hu, B., Shi, C., Zhao, W.X., Yu, P.S.: Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In: Proceedings of KDD, pp. 1531–1540 (2018)
26. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
27. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS, pp. 3111–3119 (2013)
28. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The acl anthology network corpus. In: NLPIR4DL '09 Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, vol. 47, pp. 54–61 (2009)
29. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer:extraction and mining of academic social networks. In: Proceedings of KDD, pp. 990–998 (2008)
30. Jiang, X., Sun, X., Hai, Z.: Graph-based algorithms for ranking researchers: not all swans are white! Scientometrics **96**(3), 743–759 (2013)
31. Page, L.: The PageRank Citation Ranking : Bringing Order to the Web. Stanford Digital Libraries Working Paper **9**(1), 1–14 (1999)

Figure 1: The framework of ESMR. The process of ESMR consists of three parts: (1) constructing the HAN, which includes PCN, TIN, PAN and CAN; (2) learning unified Embedding representations of entities in HAN, in which the local structure information and text information is preserved; (3) predicting the future influence of papers and authors by integrating these learned representations into a multivariate random walk based Co-Ranking model.
Figure 2: Analysis of dimensions of vectors. It is the results of top-20 paper ranking and author ranking where the dimensions of papers and authors are varied from 20 to 200. It shows that the performance of ESMR slightly varies in different dimensions and it is reasonable to select 100 as their dimensional values in the following experiments.
Figure 3: Sensitivity analysis for α_{pp} and α_{aa}. It shows the effect of α_{pp} and α_{aa} on papers and authors in term of RI and NDCG for AAN dataset. For α_{pp}, it is set to 0.3 for getting the best result. For α_{aa}, 0.15 is a reasonable choice.
Figure 4: The paper results of NDCG@k and RI@k on the AAN Dataset. It shows the performance of ESMR compared with all baselines. It demonstrates that the performance of ESMR is better than all baselines with different k. A possible explanation is that ESMR uses network embedding and adds text information, global and local structural information to improve the performance of paper impact prediction.
Figure 5: The paper results of NDCG@20 and RI@20 on the AMiner Dataset. The results in fields of Artificial Intelligence (AI) and Database(DB) in 2001, 2003,2005 and 2007 are selected to evaluate the paper ranking performance. ESMR generally outperforms all baselines, which 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
Figure 6: The author results of NDCG@k and RI@k on the AAN Dataset. The performances of our ESMR are better than all other methods at various k. we can conclude that it is necessary to integrate network embedding, text information and better ranking method.
Figure 7: The author results of NDCG@20 and RI@20 on the AMiner Dataset. It can be seen that ESMR performs the best performance in the fields of AI and DB in 2002, 2004, 2006 and 2008. 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.