Attributed Collaboration Network Embedding for Academic Relationship Mining

WEI WANG, School of Software, Dalian University of Technology, China and Department of Computer and Information Science, University of Macau, China
JIAYING LIU, School of Software, Dalian University of Technology, China
TAO TANG, School of Software, Dalian University of Technology, China
SUPPAWONG TUAROB, Faculty of Information and Communication Technology, Mahidol University, Thailand
FENG XIA*, School of Engineering, IT and Physical Sciences, Federation University Australia, Australia and School of Software, Dalian University of Technology, China
ZHIGUO GONG, State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, China
IRWIN KING, Department of Computer Science and Engineering, The Chinese University of Hong Kong, China

Finding both efficient and effective quantitative representations for scholars in scientific digital libraries has been a focal point of research. The unprecedented amounts of scholarly datasets, combined with contemporary machine learning and big data techniques, have enabled intelligent and automatic profiling of scholars from this vast and ever-increasing pool of scholarly data. Meanwhile, recent advance in network embedding techniques enables us to mitigate the challenges of large scale and sparsity of academic collaboration networks. In real-world academic social networks, scholars are accompanied with various attributes or features, such as co-authorship and publication records, which result in attributed collaboration networks. It has been observed that both network topology and scholar attributes are important in academic relationship mining. However, previous studies mainly focus on network topology, whereas scholar attributes are overlooked. Moreover, the influence of different scholar attributes are unclear. To bridge this gap, in this work, we present a novel framework of **A**ttributed **C**ollaboration **N**etwork **E**mbedding (**ACNE**) for academic relationship mining. ACNE extracts four types of scholar attributes based on the proposed scholar profiling model,

*Corresponding Author

Authors' addresses: Wei Wang, School of Software, Dalian University of Technology, China, Department of Computer and Information Science, University of Macau, China, ehomewang@ieee.org; Jiaying Liu, School of Software, Dalian University of Technology, China, jiaying_liu@outlook.com; Tao Tang, School of Software, Dalian University of Technology, China, tau.tang@outlook.com; Suppawong Tuarob, Faculty of Information and Communication Technology, Mahidol University, Thailand, suppawong.tua@mahidol.edu; Feng Xia, School of Engineering, IT and Physical Sciences, Federation University Australia, Australia, School of Software, Dalian University of Technology, China, f.xia@acm.org; Zhiguo Gong, State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, China, fstzgg@um.edu.mo; Irwin King, Department of Computer Science and Engineering, The Chinese University of Hong Kong, China, king@cse.cuhk.edu.hk.

1559-1131/2020/1-ART1 $15.00
https://doi.org/10.1145/3409736

including demographics, research, influence, and sociability. ACNE can learn a low-dimensional representation of scholars considering both scholar attributes and network topology simultaneously. We demonstrate the effectiveness and potentials of ACNE in academic relationship mining by performing collaborator recommendation on two real-world datasets and the contribution and importance of each scholar attribute on scientific collaborator recommendation is investigated. Our work may shed light on academic relationship mining by taking advantage of attributed collaboration network embedding.

CCS Concepts: • **Information systems** → **Information retrieval**; **Learning to rank**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Network embedding, academic information retrieval, scientific collaboration, graph learning

ACM Reference Format:
Wei Wang, Jiaying Liu, Tao Tang, Suppawong Tuarob, Feng Xia, Zhiguo Gong, and Irwin King. 2020. Attributed Collaboration Network Embedding for Academic Relationship Mining. *ACM Trans. Web* 1, 1, Article 1 (January 2020), 21 pages. https://doi.org/10.1145/3409736

1 INTRODUCTION

In the era of scholarly big data, scholars have increasingly published research articles that are efficiently indexed and managed by digital libraries such as Google Scholar[1] and CiteseerX[2] [21, 56]. Furthermore, many digital libraries and academic search engines, i.e. DBLP (DBLP Computer Science Bibliography)[3] and APS (American Physical Society)[4], have published their datasets to advocate and advance research in the scholarly mining fields. As a result, such datasets also enable us to study the dynamics of scholars and their roles in the academic society, by analyzing academic relationships using various cutting-edge data mining techniques that have been shown to work with large-scale scholarly data. Recent study has shown that academic relationship is highly associated with co-authorship. For example, co-authorship is a key factor that infers the scientific collaboration and academic success [60]. Advisor-advisee relationships can also be inferred based on the co-authorship networks [31, 47, 49].

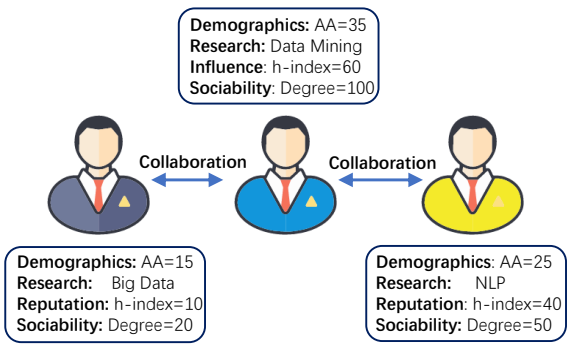


Fig. 1. Example of an attributed collaboration network.

[1]https://scholar.google.com/
[2]https://citeseerx.ist.psu.edu
[3]https://dblp.uni-trier.de/
[4]https://journals.aps.org/

Typical academic relationship mining tasks include advisor-advisee relationship identification [47, 49], collaboration pattern modelling [38], collaborator recommendation [32, 55], etc. One of the most important tasks in academic relationship mining is how to quantitatively measure the similarity between scholars. For example, most existing collaboration recommendation systems are designed based on the assumption that similar scholars are more likely to collaborate with each other [4]. Hence, it is imperative to find a scholar similarity scheme that not only is accurate, but also is appropriate to the tasks at hand.

Since research collaboration relationship can be well represented as a network of co-authors, such a network has been extensively used in many scholarly mining tasks that involve collaboration prediction. Various network-based methods for similarity measurement such as common neighbors (CN), Katz, and random walk with restart are proposed [34]. However, scholars' relationship should not be represented by only co-authorship, but also research expertise represented by their academic papers. Hence, scholars should be represented by various attributed that both reflect their existing collaboration structure and academic expertise. In this paper, such attributed networks are referred to as *attributed collaboration networks* (see Fig. 1). Regardless of the importance of scholar attributes, few attempts have been done to explore attributed collaboration networks for academic relationship mining. Although several studies have considered some academic factors, such as academic age [51] and publication content [45], these factors are incomplete and unsystematic. Moreover, the influence of different scholar attributes are unclear.

Generating attributed collaboration networks is challenging because of the nature of scholarly datasets, that are not only increasingly massive, but also composed of various information types, both structured and non-structured. Based on the results of previous study [20], at least 114 million English-language scholarly documents are accessible on the web. Second, collaboration networks are very sparse because scholars usually have a limited number of collaborators compared to the number of total scholars. Recent advances in network embedding (or network representation learning) [10] enable us to better face these two challenges by encoding network topology into a low-dimensional space. As a result, similar nodes in the embedding metrics are close to each other. The benefit and effectiveness of network embedding have been proven in many tasks, such as node classification, link prediction, and community detection [2, 5, 26, 61]. Although several attributed network embedding approaches have been proposed [16, 18, 27, 29, 48], none have been applied to scientific collaboration networks. Furthermore, to the best of our knowledge, selection of scholar attributes to represent research expertise for each node in the collaboration network has never been thoroughly investigated.

Against this background, in this work, we propose an **A**ttributed **C**ollaboration **N**etwork **E**mbedding framework for academic relationship mining named **ACNE**. ACNE can exploit both the network topology and scholar attributes simultaneously for network representation learning. Meanwhile, in order to select scholar attributes, we propose to profile scholars from four perspectives, including demographics, research, influence, and sociability. Through this, we can better depict a given scholar. Extensive experimental results on collaborator recommendation with two scholarly datasets demonstrate the effectiveness of ACNE in academic relationship mining, compared to cutting-edge baseline solutions.

The contributions of this paper can be summarized as follows:

- We propose a novel and comprehensive scholar profiling model by considering four types of scholar attribute, including demographics, research, influence, and sociability.

- We present a generic framework ACNE for academic relationship mining, which can embed collaboration networks by preserving scholar attribute and network topology.
- We conduct extensive experiments on the task of collaborator recommendation with two real-world scholarly datasets to demonstrate the effectiveness of ACNE by comparison with six state-of-the-art methods.
- We investigate the contribution and importance of each scholar attribute on scientific collaborator recommendation and find that research attribute is the most important attribute whereas demographic attribute is of lowest importance.

The rest of this paper is organized as follows. Section 2 reviews the related work. We formulate the investigated problem in Section 3. The details of ACNE are presented in Section 4. The experimental results are presented and discussed in Section 5. Finally, Section 6 concludes this paper with future work.

2 RELATED WORK

Literature on large-scale scholarly mining is extensive. In this section, we review the related work that is directly related to our problem, including network representation learning and academic relationship mining.

2.1 Network Representation Learning

Recently, approaches to embed network-like information have been extensively investigated in terms of both efficacy and generalizability [2, 3, 61]. The goal of network embedding is to learn low-dimensional vector representations of nodes in a network. Generally, such an approach can automatically generate the topology representation of nodes that encode the proximity and linkage information, using machine learning methods. Many network embedding models have been proven effective in many network-based tasks such as entity recommendation [37], link prediction [42], node classification [39], and community detection [19].

Different network representation learning approaches have been designed from various angles in terms of different network types. Notable examples of network types include heterogeneous networks [40, 52], signed networks [22, 33], attributed networks [15, 27, 29, 48, 57], and dynamic networks [9, 12]. Dong et al. [6] perform meta-path based random walks to construct the heterogeneous neighborhood of a node and then leverage a heterogeneous skip-gram model to perform heterogeneous network embedding. Yuan et al. [59] adopt the log-bilinear model to consider both the edge sign information and representations of all nodes that form a given path for singed network representation learning named SNE. Hong et al. [15] propose a deep attributed network embedding framework which adopts a personalized random walk model to calculate the relationship between network structure and node attributes using various degrees of proximity. As a result, the proposed model can capture both the network structure and attribute information synchronously. Goyal et al. [12] design a dynamic graph representation learning model which can learn the temporal transitions occurring inside the network, using a deep architecture composed of dense and recurrent layers. Although these methods are effective in learning network representations, it is still challenge to apply them to practical real-world tasks related to network dataset, i.e., academic relationship mining with scholarly big data.

In many social networks, besides relationship information represented by different weights and types of edges, each node may encode various attributes or features to quantitatively represent itself. Therefore, some researchers have paid attention to attributed network embedding [1, 11, 17, 27, 29, 35, 48, 54, 57, 62]. For example, Liao et at. [29] propose the

SNE model which can embed social networks by utilizing node attributes. Huang et al. [16] come up with the AANE model, specifically designed for learning large attributed network embedding in the distributed manner. Tu et al. [45] propose CANE model that is capable of embedding networks by considering each node's context. While network embedding approaches have been well studied, their applications in scholarly mining have been limited. In this work, our proposed method learns scholar vectors for academic relationship mining based on attributed collaboration network representation that combined both the scholar attributes and collaboration network topology.

2.2 Academic Relationship Mining

Academic relationship mining has received extensive attention in the era of scholarly big data [21, 53]. Tasks such as modeling collaboration patterns [38, 51], academic relationship identification [47], community detection [58], and collaborator recommendation [24, 32, 50, 55] have extensively been investigated. For example, Wang et al. investigate scientific collaboration patterns from scholars' local perspectives based on their academic ages [50]. They find that there is an obvious homophily phenomenon in scientific collaborations. Wang et al. [47] propose a time-constrained probabilistic-based graph model for advisor-advisee relationship identification. Yu et al. [58] propose to formulate an academic team by a barrel composed of planks, namely, Liebig's barrel. Liu et al. [32] design a context-aware collaborator recommendation framework which is consisted of two fundamental components: the collaborative entity embedding network and the hierarchical factorization model.

In this work, we mainly consider the task of collaborator recommendation. Usually, scholars explore scientific collaborations from a network perspective where two individual scholars are connected if they have coauthored at least one paper. With a collaboration network, scholar similarity can be measured via social network indices, such as common neighbors or random walk-based node similarity. Based on this idea, collaborator recommendation systems have been designed [55]. However, such methods are time-consuming and biased because these indices are manually designed and calculated. Recently, Chen et al. propose CollabSeer, a search engine for research collaborators [4]. Their system generates the co-authorship network from scientific publications, which is then used to predict infer potential collaborators using node-based similarity. Wang et al. develop a content-based method to recommend scientific articles [46]. Their method combines topical knowledge extracted from scientific papers with collaborative filtering approaches. Such a method, however, does not recommend collaborators.

Most content-based approaches take advantages of topic models, e.g., Latent Dirichlet Allocation (LDA) to calculate the topic distribution of scholars based on their paper content, such as titles and abstracts [24]. With the topic distribution, the similarity between scholars can be calculated via cosine similarity. This approach is limited in the need of correct and large-scale paper content. Meanwhile, the direct coauthorships are overlooked, which have been proven useful in academic relationship mining [46]. We believe that it is necessary to consider both scholar attributes and collaboration network topology in academic relationship mining, e.g., collaborator recommendations. To combine both pieces of information, certain issues must first be properly addressed. For example, what kind of scholar attributes should be considered? Another issue would be how to simultaneously model network topology and scholar attributes. Therefore, we aim to design an attributed-aware network representation learning model for academic relationship mining.

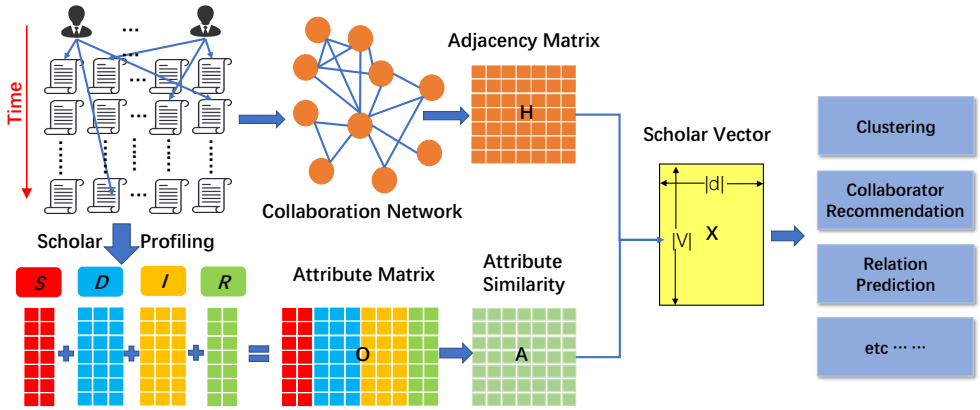


Fig. 2. Illustration of ACNE. ACNE first extracts scholars' attributes from various perspectives based on scholar profiling. Then, an attribute similarity matrix can be gained. Finally, ACNE embeds the adjacency matrix and attributed similarity matrix jointly for scholar representations.

3 PROBLEM FORMULATION

Academic relationships are more than links where scholars occupy various attributes. There are mainly two challenges in exploiting attributed collaboration networks. First, how to extract scholar attributes and which attributes would be beneficial? Second, how to project the attribute factors into a low-dimensional vector space for learning final representations. We aim at learning a low-dimensional embedding $v \in \mathbb{R}^d$ for each scholar v_i according to both the network topology and attribute information. The d in the represented space is much smaller than $|V|$, which can be used for academic relationship mining efficiently. For the rest of this work, we use boldface lowercase alphabets (e.g., $\mathbf{r}$) to present vectors and boldface uppercase $\mathbf{R}$ to present matrices. The i^{th} row of a matrix $\mathbf{R}$ is presented as $\mathbf{r}_i$. The transpose of $\mathbf{R}$ is $\mathbf{R}^T$. We denote the ℓ_2 norm of a vector using $\|\cdot\|_2$, which is the Euclidean norm of a vector.

Given a scholar in a specific digital library, e.g., DBLP, we can extract an attributed collaboration network $L = \{V, E, W, O\}$, where V is a set of scholars, O denotes the attribute information of scholars, and each edge $e_{i,j} \in E$ represents the relationship between two scholars (i, j), with an associated weight $w_{i,j}$. The weight denotes the number of collaboration times. Here, the attributes refer to the scholar characteristics based on scholar profiling, which are hidden in the raw data of the digital library.

The main symbols are shown in Table 1. In this table, the variable z_i refers to the number of column of an attribute matrix. Let $\mathbf{H} \in \mathbb{R}^{|V|\times|V|}$ be the adjacency matrix of collaboration networks. Let $\mathbf{O}$ be a $|V| \times z$ matrix that stands for all the scholar attributes where each row o_i denotes the attributes of scholar v_i. Based on the notions above, we can formulate the investigated problems as follows.

Given a set of scholars in network L with scholar attributes $\mathbf{O}$ and adjacency matrix $\mathbf{H}$, our goal is to learn a low-dimensional representation matrix $\mathbf{X}$ by preserving both scholar attributes and network topology, where each scholar $i \in V$ is represented as a low-dimensional vector $\mathbf{x}_i$. The scholar representation vectors can achieve better performance in academic relationship mining.

Table 1. Description of key symbols.

Symbols	Definitions
$n = \|V\|$	Number of scholars
w	Weight of link
e	Edge
d	Dimension of embedding result
$\mathbf{D} \in \mathbb{R}^{\|V\| \times z_1}$	Demographical attribute matrix
$\mathbf{I} \in \mathbb{R}^{\|V\| \times z_3}$	Influence attribute matrix
$\mathbf{S} \in \mathbb{R}^{\|V\| \times z_4}$	Sociability attribute matrix
$\mathbf{O} \in \mathbb{R}^{\|V\| \times z}$	Scholar attribute matrix
$\mathbf{A} \in \mathbb{R}^{\|V\| \times \|V\|}$	Attribute similarity matrix
$\mathbf{H} \in \mathbb{R}^{\|V\| \times \|V\|}$	Adjacency matrix
$\mathbf{X} \in \mathbb{R}^{\|V\| \times d}$	Final embedding matrix

4 DESIGN OF ACNE

Our ultimate goal is to learn a low-dimensional vector representation of scholars considering both collaboration network topology and scholar attributes. During the process of representations, we need to satisfy the following requirements: 1) it should utilize suitable scholar attributes in a comprehensive way; 2) it should preserve the scholar proximity both in network and attribute space well; 2) it needs to be able to handle weighted networks since the scientific collaboration networks are weighted; 4) it is necessary to be scalable since the number of scholars may be large. For these reasons, we propose an **A**ttributed **C**ollaboration **N**etwork **E**mbedding (**ACNE**) model based on matrix factorization for academic relationship mining.

The basic framework of ACNE is shown in Fig. 2. ACNE contains three parts, i.e., 1) scholar attribute extraction via scholar profiling, 2) collaboration network structure embedding, and 3) scholar attribute proximity embedding. As shown in this figure, ACNE first extracts scholar's attributes from the perspectives of demographics, research, influence, and sociability to gain an attribute matrix $\mathbf{O}$ based on the raw data of the digital library. Such a process is also called scholar profiling [43]. With the attribute matrix, we can calculate the attribute similarity matrix $\mathbf{A}$ based on cosine similarity. Then, ACNE decomposes the attribute similarity into the final scholar vector matrix $\mathbf{X}$, where the process is controlled by an edge-base penalty. Such penalty is determined by the adjacency matrix $\mathbf{H}$ of the collaboration network. Inspired by previous study [45], we can generate the scholar vector based on a joint representation learning process, where the objective of ACNE is:

$$\zeta = \zeta_A + \zeta_H, \tag{1}$$

where ζ_A denotes the objective of attribute embedding and ζ_H denotes the objective of collaboration network embedding. We will present each step in detail.

4.1 Scholar Attribute Extractions via Scholar Profiling

Scholars have various attributes so that scientific collaboration networks are attributed. The same as the network topology, scholar attributes obviously have a great influence on academic relationship mining. Previous researches have shown that scientific collaboration patterns are varied with scholar demographic characteristics, i.e., academic age [51]. The homogeny phenomenon has been found that senior scholars stick together. Thus, we need to consider

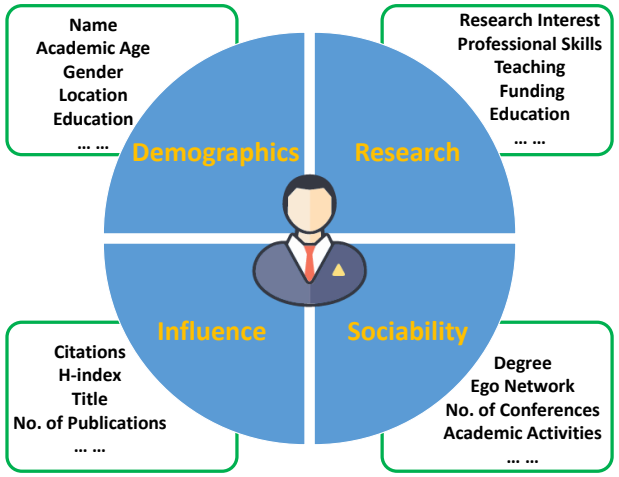


Fig. 3. Example of scholars' various kinds of attributes via scholar profiling.

scholars' attributes for academic information mining tasks. However, previous studies mainly consider network topology in tasks such as collaborator recommendations [55]. Although some researchers have utilized scholars' attributes for academic relationship mining [24], the attributes they considered are incomplete and unsystematic. The most common attribute is the research topic which is either calculated by the bag-of-words model with test information or extracted directly via tags [30, 32, 45]. Many other underlying attributes are overlooked.

In order to acquire scholar attributes in a comprehensive way, we propose to use the user profiling method [36]. User profiling aims at extracting and analyzing users' detailed characteristics for accuracy recommendation service in online social networks and e-commerce platforms. Similarly, scholars profiling aims at extracting and utilizing scholars' various attributes for academic relationship mining tasks such as citation mining and collaborator recommendation [25].

Specifically, our proposed model of scholar profiling is composed of four perspectives including, Demographics, Research, Influence, and Sociability, as shown in Fig. 3. We believe that these four types of scholars' attributes can profile scholars so that more accurate results can be obtained, which will be shown in the experiments later. The details of these four types of attributes are described as follows:

- **Demographics:** It has been proven that users with demographic profiles in social networks bring new insights into understanding social principles from individuals, to groups, and to societies [7]. The demographic characteristics of scholars are of various kinds. Some popular demographics are gender, academic age, location, nationality, etc.
- **Research:** The research attributes denote the information related to scholar's studies. This kind of scholar attribute is widely considered in previous works as the research topics are provided by some popular scholarly datasets. Meanwhile, the advances of natural language processing enable us to extract scholar's research topic distribution via topic models, e.g., LDA [24].
- **Influence:** The influence attribute refers to the indicators denoting a scholar's academic achievement and impact. This important attribute is always overlooked in academic relationship mining tasks. In reality, junior scholars are more likely to be pursuers while

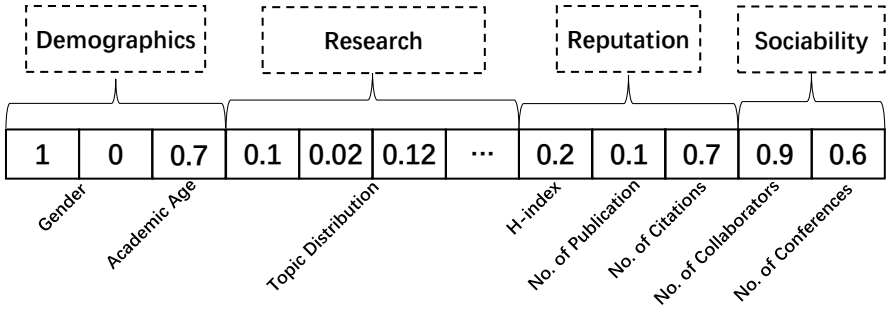


Fig. 4. Example of three kinds of scholar attributes.

senior scholars with high academic reputation are normally attractors when facing new collaborative opportunities. The influence attribute includes citations, number of publication, h-index, academic title, etc.

- **Sociability:** The sociability attribute refers to the collaboration patterns of scholars. In modern academia, some scholars are more collaborative than others. It has been proven in many studies that collaborative scholars are more productive and influential [60]. On the one hand, some network indicators can reflect the sociability attribute such as degree, ego network, and clustering coefficients. On the other hand, scholars' sociability is hidden in involving academic activities such as conference attending.

These mentioned attributes together with the co-author relationships bring about the attributed collaboration networks. It is worth mentioning that although employing all the attributes might bring better model performance, the limitations of the scholarly datasets result in the absence of some attributes. For example, gender information cannot be easily gained. Therefore, the attribute extractions are subject to the limitations exist in scholarly datasets.

4.2 Attribute-based Embedding

Scholars are associated with various attributes. Previous studies have proven that attribute information can improve the performance of network-based analysis [46]. After scholar profiling in previous section, we can obtain four types of scholar attributes, including **Demographics** (D), **Research** (R) **Influence** (I), and **Sociability** (S). Each type of attributes may contain various factors. We convert all scholar attributes into a generic feature vector, where each factor is normalized into $[0, 1]$ via Min-Max normalization method. Considering the data type, scholar attributes can be categorized into three types:

- **Isolated attributes:** Most scholar attributes are isolated, such as scholar influence attributes like the number of publications and h-index. We convert all these scholar attributes into $[0, 1]$ via Min-Max normalization.
- **Discrete attributes:** Typical discrete attributes are categorical variables, such as gender. For such attributes, we convert them into a set of binary feature. For example, a scholar's gender attribute is either *female* or *male* so that we can use the vector $\{1, 0\}$ to express a female scholar.
- **Continuous attributes:** Some scholar attributes are continuous after transformation. For example, scholars' research attributes are hidden in their publications which are

text information. We can extract scholars' topic distribution vector with the topic model.

Suppose z_1, z_2, z_3, and z_4 are the numbers of scholar's demographics, research, influence, and sociability, respectively. Thus, a scholar's attribute vector contains z features, where z is the sum of z_1, z_2, z_3, and z_4. Therefore, we can gain the scholar attribute matrix as:

$$\mathbf{O} = [\mathbf{D} + \mathbf{R} + \mathbf{I} + \mathbf{S}]. \tag{2}$$

Based on the scholar attribute matrix $\mathbf{O}$, we can gain the attribute similarity matrix $\mathbf{A}$ based on cosine similarity, where each $a_{i,j} \in \mathbf{A}$ is calculated as,

$$a_{ij} = \sum_{i=1,j=1}^{z} \mathbf{o}_i \times \mathbf{o}_j. \tag{3}$$

After the attribute similarity matrix is computed, to preserve the attribute proximity, we propose the optimize objective by approximating the scholar attribute similarity matrix $\mathbf{A}$ with the final scholar vector matrix $\mathbf{X}$, which can be calculated as:

$$\zeta_A = \| \mathbf{A} - \mathbf{X}\mathbf{X}^T \|_F^2 = \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} (\mathbf{a}_{ij} - \mathbf{x}_i \mathbf{x}_j^T)^2. \tag{4}$$

The goal of this objective function is to minimize the difference between dot product of vector representation $\mathbf{x}_i$ and $\mathbf{x}_j$, and the corresponding paper attribute similarity $\mathbf{a}_{ij}$.

4.3 Structure-based Embedding

The structure-based embedding aims to measure the log-likelihood of an edge. Following the DeepWalk model [42], the objective can be calculated as:

$$\zeta_H = w_{i,j} \log p(v_i|v_j), \tag{5}$$

where $\log p(v_i|v_j)$ denotes the conditional probability of v_i generated by v_j, which can be calculated as:

$$\log p(v_i|v_j) = \frac{\exp(v_i \cdot v_j)}{\sum_{z \in V} \exp(v_i \cdot v_j)}. \tag{6}$$

However, this loss function cannot be easily learned via matrix factorization. Following previous work [16], without loss of generality and for ease of calculation, we use the following loss function instead of Eq. (5):

$$\zeta_H = \sum_{(i,j) \in e} w_{ij} ||\mathbf{x}_i - \mathbf{x}_j||_2, \tag{7}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the representations of scholars i and j, w_{ij} denotes the links weight between them. It can measure the log-likelihood of an edge between x_i and x_j via matrix factorization by calculating the differences between two scholars. The key motivation is that in order to minimize the penalty $w_{ij}||\mathbf{x}_i - \mathbf{x}_j||$, a larger w_{ij} may lead to a smaller difference between $\mathbf{x}_i$ and $\mathbf{x}_j$.

4.4 Joint Representation Learning

Based on Eq. (4) and Eq.(7), we can rewrite the objective of ACNE as:

$$\min_{\mathbf{X}} \zeta = \| \mathbf{A} - \mathbf{X}\mathbf{X}^T \|_F^2 + \delta \sum_{(i,j) \in e} w_{ij} ||\mathbf{x}_i - \mathbf{x}_j||_2, \tag{8}$$

where δ denotes a trade-off between scholar attributes and collaboration network. This objective function can jointly model the network topology and scholar attributes. Since this objective function is bi-convex [16], we adopt the accelerated and distributed algorithm for parameter optimization in AANE. The key idea is to accelerate the optimization by converting it into 2n updating steps and one matrix updating step.

5 EXPERIMENTS

In order to investigate the potentials of ACNE in academic relationship mining, we conduct experiments of collaborator recommendation on two scholarly datasets DBLP and APS. Collaborator recommendation has been an extensively studied task, and many advanced recommendation strategies have been proposed. Specifically, we try to answer the following research questions.

- **RQ1** Is it beneficial to consider scholar attributes for academic relationship mining?
- **RQ2** Which kind of scholar attribute is more useful?
- **RQ3** Can ACNE better represent scholars for collaborator recommendation as compared to state-of-the-art methods?

5.1 Experimental Design

The experimental designs including evaluation metrics, comparison methods, and parameter tuning are introduced in this section.

5.1.1 Dataset and Setup. We adopt two widely used scholarly datasets, i.e., DBLP and APS datasets. Specifically, the Aminer [5] scholarly platform [44] has provided the preprocessed DBLP dataset which can be downloaded online [6]. The APS dataset can be freely accessed online [7] upon request. For each dataset, we first perform name disambiguation using the method in [41]. Then, we filter out those scholars who have an academic career less than 5 years or with less than 10 publications [41]. For the DBLP datasets, we gain the citation relationships from AMiner project [44]. In this work, we utilize a subset of DBLP in the field of data mining, and a subset of APS with the journals of Physic Review A, B, and C. The statistics of two datasets are shown in Table 2.

Due to the limitation of these two datasets, we cannot gain all the proposed scholar attribute. The demographical attribute we use is the academic age which is calculated by the investigated year minus the year of first publication. This demographic characteristic has been proven influential in collaboration patterns [51]. The research attributes are the topic distributions which are calculated by performing topics model LDA on the title and abstract information. The influence attributes include number of citations, number of publications, and h-index. The sociability attributes adopted are numbers of collaborators and clustering coefficient [28], where the clustering coefficient indicates how close you and your neighbors are.

Without loss of generality, during the experiments, we set the number of generated topics by LDA as 100. The Top-k candidates are ranked by the cosine similarity with two scholar vectors in $\mathbf{X}$. It is worth mentioning that we aim at recommending new collaborators. Therefore, the Top-k candidates are selected by filtering out those scholars who had collaborated with target scholar before. For collaborator recommendations, we split the datasets into two subsets, including the training set with $t = [2008, 2010]$ and the testing set with $t = [2011, 2015]$. If

[5]https://aminer.org/
[6]https://aminer.org/billboard
[7]https://journals.aps.org/datasets

Table 2. Statistics of two datasets.

Datasets	#Scholars	#Papers	#Links
DBLP	59,659	21,675	90,282
APS	3,784	1,218	8,395

two unconnected scholars in the training set coauthor with each other at least one paper in the testing set, we regard it as a positive sample. We use 100% fractions of data in training set for parameter tuning.

5.1.2 Evaluation Metrics. We adopt three widely used metrics for recommendation system evaluation, including Precison@k, Recall@k, and F1@k. These metrics have been extensively used to evaluate the performance of a recommendation system.

The Precision@k is defined as:

$$Precision@k = \frac{\#\ of\ new\ collaborators\ in\ the\ Top-k\ list}{\#\ of\ k}. \tag{9}$$

The Recall@k is defined as:

$$Recall@k = \frac{\#\ of\ new\ collaborators\ in\ the\ Top-k\ list}{total\ \#\ of\ new\ collaborators}. \tag{10}$$

The F1@k is an integrated index of Precision@k and Recall@k, which is defined as:

$$F1@k = \frac{2 \times Precision@k \times Recall@k}{Precision@k + Recall@k}. \tag{11}$$

5.1.3 Baseline Methods. We compare ACNE with two categories of baselines or state-of-the-art collaborator recommendation approaches. In order to evaluate the effectiveness of considering scholar attributes, two plain network embedding models, and a context-aware network embedding model are used, i.e., Node2vec [13], DeepWalk [37], and CANE [45]. Meanwhile, three typical collaborator recommendation methods are used for comparison, i.e., Common Neighbor (CN), TopicSim, and MVCWalker [55]. The details of these methods are listed as follows.

- **CN:** CN is a network-based similarity measurement, which can be used to recommend collaborators based on the number of common neighbors. It is very simple and has been extensively used in link prediction and collaborator recommendation.
- **TopicSim:** TopicSim recommends collaborators based on the topic similarity between scholars. The topic vectors are calculated by performing topic model, i.e., LDA on scholars' publication text.
- **MVCWalker:** MVCWalker [55] is a random-walk-based collaborator recommendation approach which considers three factors to calculate link weights, i.e., coauthor order, latest collaboration time point, and collaboration times.
- **Node2vec:** Node2vec [13] adopts the Skip-Gram model on node sequence in collaboration networks generated by biased random walk. The gained scholar vector can be used for similarity calculation for recommendation.
- **DeepWalk:** DeepWalk [37] utilizes word2vec and truncated random walk techniques for network embedding. Both Node2vec and DeepWalk are plain network embedding approaches which do not consider any scholar attribute.
- **GCN:** Graph Convolution Network [23] is a new method for processing structured data with deep learning. This method is mainly based on the representation and

Table 3. Parameter sensitivity of $|d|$. k is set as 10.

d	DBLP (%)			APS (%)		
	P@k	R@k	F1@k	P@k	R@k	F1@k
10	28.25	18.21	22.14	37.14	22.14	27.74
20	28.73	18.35	22.39	37.35	22.32	27.94
30	29.02	18.67	22.72	**37.42**	**22.38**	**28.01**
40	**29.21**	**18.97**	**23.01**	37.41	22.24	27.89
50	29.17	18.83	22.89	37.32	22.18	27.82

calculation of the structured neural network model in the form of a graph. We adopt the model of GraphSage [14] as the GCN method.

- **DANE:** Deep Attributed Network Embedding [11] is a novel deep attributed network embedding approach, which can capture the high nonlinearity and preserve various proximities in both topological structure and node attributes. During experiments, we adopt the research interest as the attribute in DANE.
- **CANE:** CANE [45] is a context-aware network embedding approach which can learn to model the semantic relationships between nodes. In the experiments, the context information is extracted from scholars' publication records, including paper metadata such as titles and abstracts.

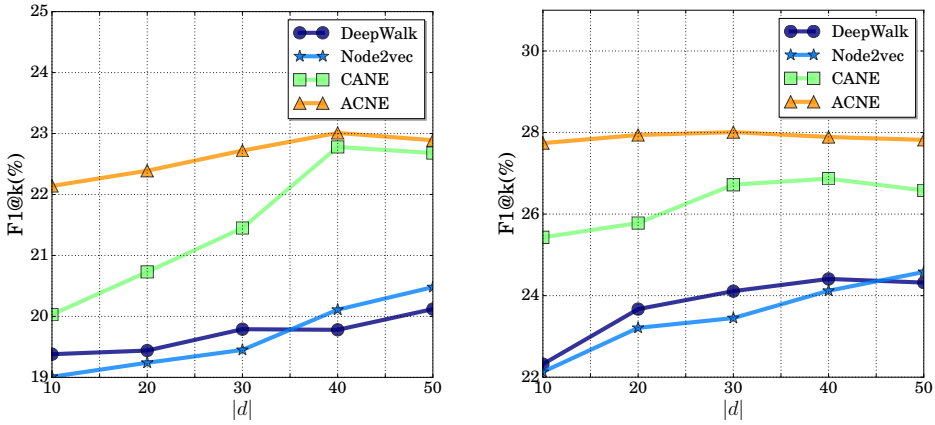
5.1.4 Parameter Sensitivity. Parameter $|d|$ denotes the dimension size of represented matrix $\mathbf{X}$. This parameter is shared across the experiments of network embedding approaches. During experiments, $|d|$ is ranged from {10, 20, 30, 40, 50}. We test the parameter sensitivity of $|d|$ by setting k as 10 during recommendation.

Table 3 shows the results of Precision@k, Recall@k, and F1@k with different $|d|$ in DBLP and APS datasets. We can observe that the performance of ACNE on DBLP tends to saturate when $|d|$ is 40, while the performance on APS tends to saturate when $|d|$ reaches 30. In the rest experiments, we set $|d|$ as 40 and 30 for DBLP and APS, respectively. At the same time, we can observe that the performance of ACNE on APS dataset is much better than that of DBLP dataset. The reason may be that the network constructed from APS dataset is denser than that of DBLP dataset.

5.2 Benefit of Attributes (RQ1)

To illustrate the benefit of considering scholar attributes in academic relationship mining. We vary the embedding dimension $|d|$ and compare the proposed ACNE model with other plain network embedding models without considering scholar attributes. These methods include Node2vec, DeepWalk, and CANE. Fig. 5 shows the F1@k scores of each method. The key observations are as follows:

(1) The proposed ACNE model achieves the best performance in terms of F1@k among all network embedding methods. It can be notably observed that, compared to the pure structure-based methods, i.e., Node2vec and DeepWalk, our proposed ACNE achieves a significantly better performance. This can be seen as a piece of evidence that considering scholar attributes is beneficial in academic relationship mining, especially those relying on collaboration recommendation. Moreover, we can observe that the ACNE model is more stable than other network embedding methods when using a smaller $|d|$.



(a) DBLP (b) APS

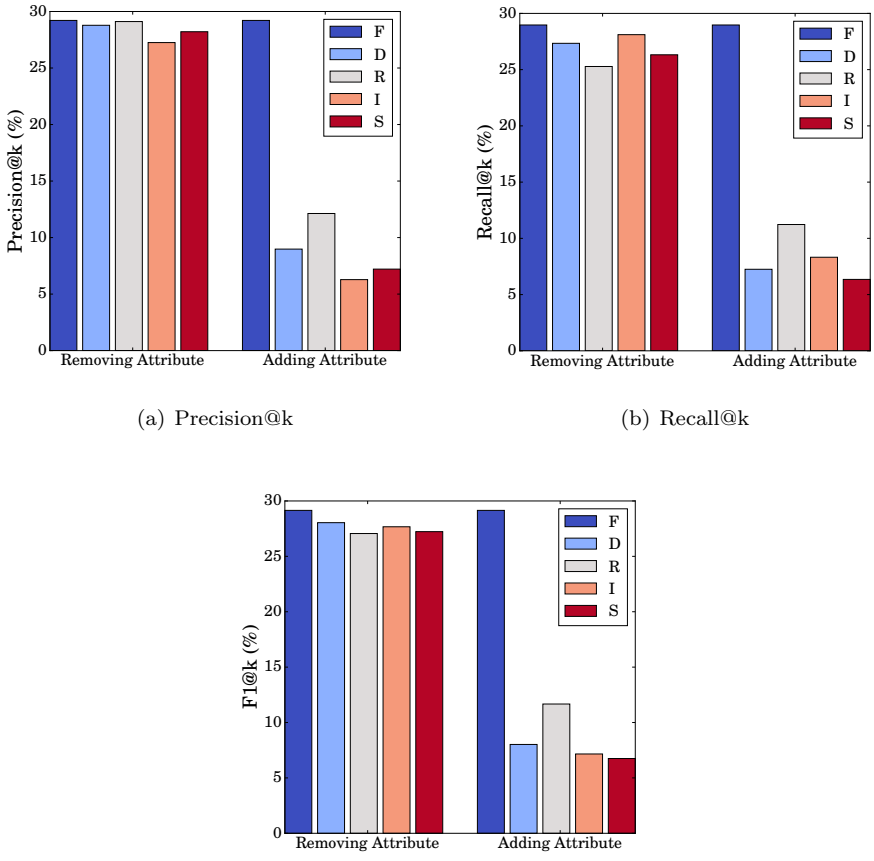
Fig. 5. Comparison between ACNE with other network embedding models in terms of F1@k over representation dimension $|d|$. We set k as 10.

(2) For plain network embedding approaches, CANE outperforms Node2vec and DeepWalk and is second only to ACNE. The reason is that CANE considers scholars' context information which is extracted from scholars' publication records. Such context information, to some extent, can be regarded as the research attribute. This observation further demonstrates the usefulness of considering scholar attributes.
(3) Comparing the performances on two different datasets, we can observe that all methods have a better recommendation accuracy on APS than that on DBLP. The reason is that the collaboration network of APS is denser than DBLP. Meanwhile, scholars in APS, whose majority of papers are in physics fields, collaborate more frequently than scholars in DBLP. This stronger collaborative behaviors observed in among APS scholars may strengthen collaborative signals in the learned models.

5.3 Attribute Analysis (RQ2)

Based on scholar profiling, we have extracted four types of scholar attributes. Some scholar attributes may have a more significant impact on academic relationship mining. To explore the contribution and importance of each type of scholar attributes, we adopt the "Jackknife" [8] approach, which contains two kinds of strategies, i.e., 1) Removing one type of scholar attribute and embedding with the rest attribute (Removing Attribute); 2) Using only one type of scholar attribute for network embedding (Adding Attribute). This method can investigate the individual contribution, and unique information that each type of scholar attribute supplies to the overall recommendation task.

Figures 6 and 7 show the recommendation performance of ACNE with "Jackknife" approach on DBLP and APS datasets, respectively. In this experiment, we merely use the ACNE model with $|d|$ as 40, 30 for DBLP and APS, respectively. The number of recommended candidates k is set as 10. The key observations of these two group figures are as follows:



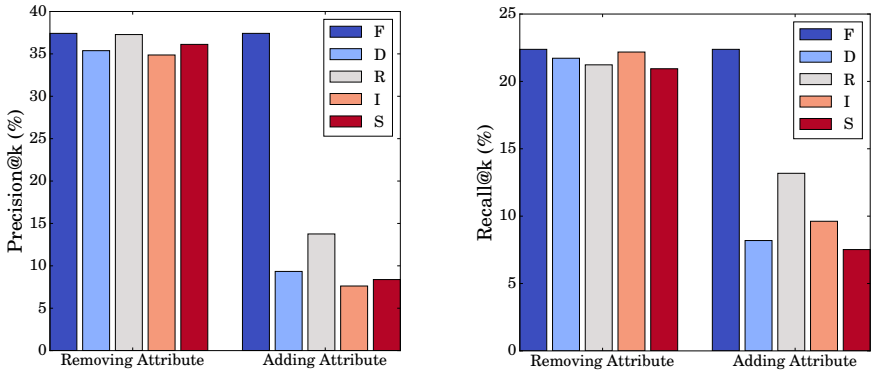
(a) Precision@k

(b) Recall@k

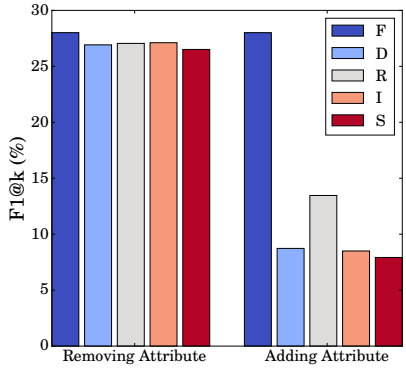
(c) F1@k

Fig. 6. Attribute contribution analysis in DBLP. F : Full attribute set; D : Demographic attribute set; R : Research attribute set; I : Influence attribute set; S : Sociality attribute set; $|d| = 30$; $k = 10$.

(1) Considering all four types of scholar attributes (Full attribute set) achieves the best performance among all strategies in terms of Precison@k, Recall@k, and F1@k on both datasets. Meanwhile, the removing strategy regarding each type of scholar attribute has a smaller but similar performance compared to using full attribute set. These demonstrate the effectiveness of our proposed scholar profiling approaches.
(2) Based on the performance of adding strategy of each type of scholar attribute, we can observe that there is a dramatic performance drop compared to removing strategy. This indicates that merely adopting a single type of scholar attribute will lead to low recommendation accuracy. Meanwhile, all the strategies have a better performance on APS than that on DBLP.
(3) Focusing on the performance of adding strategy of each type of scholar attribute, we can see that the research attribute always has a highest F1@k score than other scholar attributes and the demographic attribute has the lowest score. We can infer



(a) Precision@k (b) Recall@k



(c) F1@k

Fig. 7. Attribute contribution analysis in APS. F : Full attribute set; D : Demographic attribute set; R : Research attribute set; I : Influence attribute set; S : Sociality attribute set; $|d| = 30$; $k = 10$.

that research attribute is the most important attribute whereas demographic attribute is of lowest importance. In fact, adding strategy with research attribute is, to some extent, the same as CANE model because they both integrate scholars' publication information into network embedding.

(4) Another interesting observation is that just using the topological structure is better than the adding strategy. The possible reason is that merely considering one typo of attributes can not profile a scholar precisely, which may lead to biased results. The reason why two scholars collaborate with each other is unclear. Thus, merely consider one type of scholar attribute, i.e., the academic age attribute may hurt the performance of merely using topology structure. This also indicates that the topology structure is more important for collaborator recommendation. This is why most previous collaborator recommendation models are designed based on network

Table 4. Comparison between ACNE and baseline models in terms of Precision@k, Recall@k, and F1@k in DBLP. We set $|d|$ as 40 for every network embedding model and $k \in \{5, 10, 15, 20\}$.

DBLP	Precision@k(%)				Recall@k(%)				F1@k(%)			
	5	10	15	20	5	10	15	20	5	10	15	20
CN	22.21	19.32	13.24	10.79	14.45	15.34	16.77	17.26	17.94	17.1	14.8	13.29
TopicSim	18.44	15.22	10.21	8.76	13.23	14.23	15.11	16.74	17.51	14.71	12.19	11.5
MVCWalker	26.34	25.33	12.18	10.38	16.46	17.98	19.13	20.08	15.41	21.03	14.88	13.69
DeepWalk	25.45	23.92	15.41	11.38	15.47	17.29	18.11	18.36	20.26	20.07	16.65	14.05
Node2vec	25.41	24.81	14.46	12.39	15.24	16.37	18.21	19.54	19.24	19.73	16.12	15.16
GCN	26.73	25.14	20.35	13.51	15.85	16.42	18.71	19.72	19.37	20.18	19.87	16.89
CANE	29.38	22.11	**21.36**	15.22	17.05	19.26	19.31	19.78	19.05	20.59	20.28	17.2
DANE	30.12	**29.33**	21.22	16.02	17.22	18.34	19.08	19.89	20.44	**23.45**	19.22	18.14
ACNE	**30.28**	29.21	20.32	**16.32**	**17.18**	**18.97**	**19.33**	**20.87**	**21.58**	23.01	**19.81**	**18.32**

Table 5. Comparison between ACNE and baseline models in terms of Precision@k, Recall@k, and F1@k in APS. We set $|d|$ as 30 for every network embedding model and $k \in \{5, 10, 15, 20\}$.

APS	Precision@k(%)				Recall@k(%)				F1@k(%)			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**	**5**	**10**	**15**	**20**
CN	24.3	22.33	16.67	12.11	17.32	18.23	19.44	19.89	20.22	20.07	17.95	15.05
TopicSim	20.12	18.32	15.22	11.03	15.34	17.21	18.93	19.03	17.41	17.75	16.87	13.97
MVCWalker	38.28	33.8	26.19	13.22	20.28	21.46	24.33	28.8	26.51	26.25	25.22	18.12
DeepWalk	36.06	31.28	23.56	13.1	18.87	19.34	21.02	22.98	24.78	23.9	22.21	16.69
Node2vec	35.22	30.21	22.34	12.87	18.34	19.22	20.14	22.7	24.12	23.49	21.18	16.43
GCN	36.71	31.36	23.45	14.77	19.34	20.12	21.57	23.45	24.56	23.22	22.14	18.23
CANE	39.21	34.2	26.38	14.01	19.97	20.43	22.34	25.21	26.46	25.59	24.19	18.01
DANE	39.47	36.22	27.15	15.11	20.45	21.22	23.15	26.21	26.89	27.21	24.33	18.92
ACNE	**40.22**	**37.42**	**27.38**	**15.22**	**21.27**	**22.38**	**25.31**	**29.72**	**27.82**	**28.01**	**26.3**	**20.13**

topologies, such as random walk and collaborative filtering. This also demonstrates the significance of our proposed scholar profiling approach.

5.4 Comparison with Baselines (RQ3)

In this final subsection, we explore the potentials of ACNE in academic relationship mining based on the task of collaborator recommendation by comparison with the state-of-the-art collaborator recommendation methods.

The comparisons between ACNE and all baseline models in terms of Precision@k, Recall@k, and F1@k on both DBLP and APS datasets are illustrated in Tables 4 and 5, respectively. During the experiments, we set $|d|$ as 40 for every network embedding model and the number of recommended candidates k ranges in $\{5, 10, 15, 20\}$. The key observations based on these two tables are as follows:

(1) Our proposed ACNE can achieve the best performance over almost all evaluation metrics among all the recommendation methods. Notably, by comparison with the CN, a classical link prediction method, in terms of F1@k, ACNE has a significant improvement. Specifically, there is a 35.29% and 39.56% increase in F1@k with $k = 10$ in DBLP and APS datasets, respectively. This indicates that it is helpful to consider academic factors in designing collaborator recommendation systems, which is in line with previous studies [55].

(2) TopicSim has the worst performance among all methods, which indicates that merely considering scholar attributes without network topology is insufficient for designing a collaborator recommendation system. Meanwhile, network embedding methods, i.e.,

DeepWalk, Node2vec and GCN achieve better performance than CN, which indicates that network embedding techniques can better learn network proximity. Aligning with the findings in [4] that suggest that structural similarity between nodes could be more useful than context similarity, and that could discover authors who share similar research interests. It is worth mentioning that DANE achieves the second best performance. The reason is that DANE is, to some extent, the same as ACNE under adding strategy with scholars' interest attributes.

(3) On the one hand, with the increasing the k, the Recall@k of all recommendation approaches gradually increase. This is because of the definition of Recall@k. Based on its definition, with the increasing of k, more accurate candidates may be recommended whereas the number of new collaborators is stable. On the other hand, the Precision@k of all recommendation approaches gradually decrease with the increasing of k. This is also because of its definition. F1@k is an integrated index which can better evaluate a recommendation system. We can observe that all methods achieve the highest F1@k when k is set as 10.

6 CONCLUSION

To better understand scholars ourselves and academic society, we need to better utilize the power of scholarly big data. Previous studies on scholar similarity measurement focus either on the research topic or network topology in academic relationship mining. Academic relationships are more than links in collaboration networks because scholars occupy various academic specific attributes. A hybrid method that considers both scholar attributes and network topology is needed. To this end, we propose a generic attributed collaboration network embedding model named **ACNE**. Before performing network embedding, ACNE first extracts four types of scholar attributes based on the proposed scholar profiling model, including demographics, research, influence, and sociability. As a result, ACNE can learn a low-dimensional representation of scholars considering both scholar attributes and network topology simultaneously. Extensive results on the task of collaborator recommendation in two real-world datasets by comparison with state-of-the-art methods demonstrate the effectiveness of ACNE in academic relationship mining.

This work has tackled network embedding on scientific collaboration networks by considering both scholar attributes and network topology. We believe such a hybrid method can better measure scholar similarity so that more accurate academic relationship mining results can be gained. In future work, we will test the effectiveness of ACNE on other academic relationship mining tasks such as advisor-advisee relationship identification and community detection. Moreover, considering the fact that different features may have different weights to measure the similarity of scholars, we will explore advanced technologies such as attention mechanism to weight the importance of different feature views.

ACKNOWLEDGMENTS

This work is partially supported by National Natural Science Foundation of China under Grant No. 61872054, the Fundamental Research Funds for the Central Universities (DUT19LAB23), and the China Postdoctoral Science Foundation (2019M651115).

REFERENCES

[1] Uchenna Akujuobi, Han Yufei, Qiannan Zhang, and Xiangliang Zhang. 2019. Collaborative graph walk for semi-supervised multi-label node classification. *arXiv preprint arXiv:1910.09706* (2019).

[2] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).

[3] H. Cai, V. W. Zheng, and K. C. Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637. https://doi.org/10.1109/TKDE.2018.2807452

[4] Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, and Clyde Lee Giles. 2011. Collabseer: a search engine for collaboration discovery. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. ACM, 231–240.

[5] Yankai Chen, Jie Zhang, Yixiang Fang, Xin Cao, and Irwin King. 2020. Efficient Community Search over Large Directed Graph: An Augmented Index-based Approach. In *IJCAI*. 3544–3550.

[6] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 135–144.

[7] Yuxiao Dong, Nitesh V. Chawla, Jie Tang, Yang Yang, and Yang Yang. 2017. User Modeling on Demographic Attributes in Big Mobile Social Networks. *ACM Trans. Inf. Syst.* 35, 4, Article 35 (July 2017), 33 pages. https://doi.org/10.1145/3057278

[8] Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. 2016. Can scientific impact be predicted? *IEEE Transactions on Big Data* 2, 1 (2016), 18–30.

[9] Lun Du, Yun Wang, Guojie Song, Zhicong Lu, and Junshan Wang. 2018. Dynamic Network Embedding: An Extended Approach for Skip-gram based Network Embedding.. In *IJCAI*. 2086–2092.

[10] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. In *The Web Conference (WWW)*. Taipei, Taiwan, 2331–2341.

[11] Hongchang Gao and Heng Huang. 2018. Deep attributed network embedding. In *IJCAI*. 3364–3370.

[12] Palash Goyal, Sujit Rokka Chhetri, and Arquimedes Canedo. 2019. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning. *Knowledge-Based Systems* (2019). https://doi.org/10.1016/j.knosys.2019.06.024

[13] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.

[14] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*. 1024–1034.

[15] R. Hong, Y. He, L. Wu, Y. Ge, and X. Wu. 2019. Deep attributed network embedding by preserving structure and attribute information. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2019), 1–12. https://doi.org/10.1109/TSMC.2019.2897152

[16] Xiao Huang, Jundong Li, and Xia Hu. 2017. Accelerated attributed network embedding. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 633–641.

[17] Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. 2010. Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 570–586.

[18] Yantao Jia, Yuanzhuo Wang, Xiaolong Jin, Hailun Lin, and Xueqi Cheng. 2018. Knowledge graph embedding: A locally and temporally adaptive translation-based approach. *ACM Transactions on the Web (TWEB)* 12, 2 (2018), 8.

[19] Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. 2018. Cross-language citation recommendation via hierarchical representation learning on heterogeneous graph. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 635–644.

[20] Madian Khabsa and C Lee Giles. 2014. The number of scholarly documents on the public web. *PloS ONE* 9, 5 (2014), e93949.

[21] Samiya Khan, Xiufeng Liu, Kashish A Shakil, and Mansaf Alam. 2017. A survey on scholarly data: From big data perspective. *Information Processing & Management* 53, 4 (2017), 923–944.

[22] Junghwan Kim, Haekyu Park, Ji-Eun Lee, and U Kang. 2018. Side: representation learning in signed directed networks. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 509–518.

[23] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[24] Xiangjie Kong, Huizhen Jiang, Teshome Megersa Bekele, Wei Wang, and Zhenzhen Xu. 2017. Random walk-based beneficial collaborators recommendation exploiting dynamic research interests and academic influence. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1371–1377.

[25] Ronald N Kostoff, J Antonio del Rio, James A Humenik, Esther Ofilia Garcia, and Ana Maria Ramirez. 2001. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology* 52, 13 (2001), 1148–1156.

[26] Jianxin Li, Taotao Cai, Ke Deng, Xinjue Wang, Timos Sellis, and Feng Xia. 2020. Community-diversified Influence Maximization in Social Networks. *Information Systems* 92 (2020).

[27] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. 2017. Attributed network embedding for learning in a dynamic environment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 387–396.

[28] Yusheng Li, Yilun Shang, and Yiting Yang. 2017. Clustering coefficients of large networks. *Information Sciences* 382 (2017), 350–358.

[29] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2018. Attributed social network embedding. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (2018), 2257–2270.

[30] Han Liu, Xianchao Zhang, and Xiaotong Zhang. 2018. Possible world based consistency learning model for clustering and classifying uncertain data. *Neural Networks* 102 (2018), 48–66.

[31] Jiaying Liu, Feng Xia, Lei Wang, Bo Xu, Xiangjie Kong, Hanghang Tong, and Irwin King. 2019. Shifu2: A Network Representation Learning Based Model for Advisor-advisee Relationship Mining. *IEEE Transactions on Knowledge and Data Engineering* (2019). https://doi.org/10.1109/TKDE.2019.2946825

[32] Zheng Liu, Xing Xie, and Lei Chen. 2018. Context-aware academic collaborator recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1870–1879.

[33] Chunyu Lu, Pengfei Jiao, Hongtao Liu, Yaping Wang, Hongyan Xu, and Wenjun Wang. 2019. SSNE: Status signed network embedding. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 81–93.

[34] Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications* 390, 6 (2011), 1150–1170.

[35] Zaiqiao Meng, Shangsong Liang, Hongyan Bao, and Xiangliang Zhang. 2019. Co-embedding attributed networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 393–401.

[36] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. 2004. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 54–88.

[37] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 701–710.

[38] Alexander Michael Petersen. 2015. Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences* 112, 34 (2015), E4671–E4680.

[39] Dominic Seyler, Praveen Chandar, and Matthew Davis. 2018. An information retrieval framework for contextual suggestion based on heterogeneous information network embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 953–956.

[40] Yu Shi, Qi Zhu, Fang Guo, Chao Zhang, and Jiawei Han. 2018. Easing embedding learning by comprehensive transcription of heterogeneous information networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2190–2199.

[41] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. Quantifying the evolution of individual scientific impact. *Science* 354, 6312 (2016), aaf5239.

[42] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[43] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. 2010. A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5, 1 (2010), 2.

[44] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 990–998.

[45] Cunchao Tu, Han Liu, Zhiyuan Liu, and Maosong Sun. 2017. CANE: Context-aware network embedding for relation modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), Vol. 1. 1722–1731.

[46] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 448–456.

[47] Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. 2010. Mining advisor-advisee relationships from research publication networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 203–212.

[48] Suhang Wang, Charu Aggarwal, Jiliang Tang, and Huan Liu. 2017. Attributed signed network embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 137–146.

[49] Wei Wang, Jiaying Liu, Feng Xia, Irwin King, and Hanghang Tong. 2017. Shifu: Deep learning based advisor-advisee relationship mining in scholarly big data. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 303–310.

[50] Wei Wang, Jiaying Liu, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2019. Sustainable collaborator recommendation based on conference closure. *IEEE Transactions on Computational Social Systems* 6, 2 (2019), 311–322.

[51] Wei Wang, Shuo Yu, Teshome Megersa Bekele, Xiangjie Kong, and Feng Xia. 2017. Scientific collaboration patterns vary with scholars' academic ages. *Scientometrics* 112, 1 (2017), 329–343.

[52] Yueyang Wang, Ziheng Duan, Binbing Liao, Fei Wu, and Yueting Zhuang. 2019. Heterogeneous attributed network embedding with graph convolutional networks. *Methods* 25, 50 (2019), 75.

[53] Kyle Williams, Jian Wu, Sagnik Ray Choudhury, Madian Khabsa, and C Lee Giles. 2014. Scholarly big data information extraction and integration in the citeseer χ digital library. In *2014 IEEE 30th International Conference on Data Engineering Workshops*. IEEE, 68–73.

[54] Wei Wu, Bin Li, Ling Chen, and Chengqi Zhang. 2018. Efficient attributed etwork embedding via recursive randomized hashing.. In *IJCAI*. 2861–2867.

[55] Feng Xia, Zhen Chen, Wei Wang, Jing Li, and Laurence T Yang. 2014. Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *IEEE Transactions on Emerging Topics in Computing* 2, 3 (2014), 364–375.

[56] Feng Xia, Wei Wang, Teshome Megersa Bekele, and Huan Liu. 2017. Big scholarly data: A survey. *IEEE Transactions on Big Data* 3, 1 (2017), 18–35.

[57] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. 2015. Network representation learning with rich text information.. In *IJCAI*. 2111–2117.

[58] Shuo Yu, Feng Xia, and Huan Liu. 2019. Academic team formulation based on liebig's barrel: discovery of anticask effect. *IEEE Transactions on Computational Social Systems* (2019), 1–12. https://doi.org/10.1109/TCSS.2019.2913460

[59] Shuhan Yuan, Xintao Wu, and Yang Xiang. 2017. SNE: signed network embedding. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 183–195.

[60] Chenwei Zhang, Yi Bu, Ying Ding, and Jian Xu. 2018. Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology* 69, 1 (2018), 72–86.

[61] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2018. Network representation learning: A survey. *IEEE Transactions on Big Data* (2018). https://doi.org/10.1109/TBDATA.2018.2850013

[62] Zhen Zhang, Hongxia Yang, Jiajun Bu, Sheng Zhou, Pinggang Yu, Jianwei Zhang, Martin Ester, and Can Wang. 2018. ANRL: Attributed network representation learning via deep neural networks.. In *IJCAI*. 3155–3161.