Shifu: Deep Learning Based Advisor-advisee Relationship Mining in Scholarly Big Data

Wei Wang, Jiaying Liu, Feng Xia
School of Software, Dalian University of Technology, China
f.xia@acm.org

Irwin King
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
king@cse.cuhk.edu.hk

Hanghang Tong
School of Computing, Informatics and Decision Systems Engineering, Arizona State University, USA
hanghang.tong@asu.edu

ABSTRACT

Scholars in academia are involved in various social relationships such as advisor-advisee relationships. The analysis of such relationship can provide invaluable information for understanding the interactions among scholars as well as providing many researcher-specific applications such as advisor recommendation and academic rising star identification. However, in most cases, high quality advisor-advisee relationship dataset is unavailable. To address this problem, we propose Shifu, a deep-learning-based advisor-advisee relationship identification method which takes into account both the local properties and network characteristics. In particular, we explore how to crawl advisor-advisee pairs from PhDtree project and extract their publication information by matching them with DBLP dataset as the experimental dataset. To the best of our knowledge, no prior effort has been made to address the scientific collaboration network features for relationship identification by exploiting deep learning. Our experiments demonstrate that the proposed method outperforms other state-of-the-art machine learning methods in precision (94%). Furthermore, we apply Shifu to the entire DBLP dataset and obtain a large-scale advisor-advisee relationship dataset.

Keywords

Scholarly Big Data; Deep Learning; Relation Mining; Coauthor Network.

1. INTRODUCTION

The role of advisors in advisee future performance is adamantly important. Previous literature suggests that both advisors and advisees benefit from the advisor-advisee relationships [1, 2]. For instance, while advisees receive specialized guidance, social support, and career coaching from the advisor, advisors also benefit from the collaborations with their advisees [3]. Institutions benefit from advisor-advisee relationships as well because advisees are more likely to work for their institutions after graduation, which leads to the organizational citizenship behaviors [4]. Awareness of the advisor-advisee relationships can provide significant information for many researcher-specific applications such as advisor recommendation and academic rising star identification. For example, based on the advisor-advisee relationships, we can better understand what attributes a great advisor has, how the advisors' academic performance influences the future development of advisees, and how to predict the future success of the junior scholars.

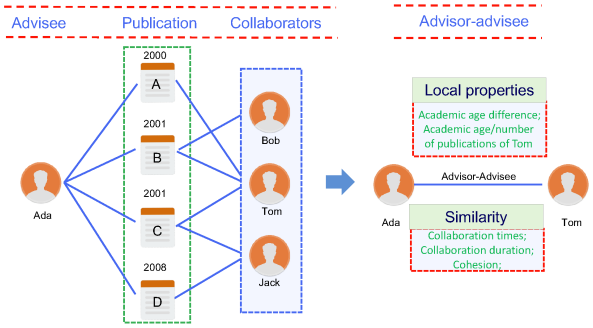


Figure 1: An example of advisor-advisee relationship in scientific collaboration networks.

However, lack of high-quality annotated dataset makes it challenging to investigate advisor-advisee relationships. Although there are several projects that aim to collect such relationships, such as Mathematics Genealogy Project[1], The Academic Family Tree[2], and PhDTree project[3], they still suffer from some drawbacks. On the one hand, these projects heavily rely on manual efforts, which results in limited records. On the other hand, most of the existing projects provide little background or contextual information, which limits the feasibility of utilizing automatic approaches, i.e., supervised learning. The anatomy of advisor-advisee relationships requires other academic information such as publications, collaborators, and h-index. An ideal solution to generate a suitable advisor-advisee dataset is to design a method that can automatically extract the relationships from scholarly big data such as DBLP (Digital Bibliography and Library Project), MAG (Microsoft Academic Search),

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054159

[1]http://genealogy.math.ndsu.nodak.edu/index.php
[2]http://academictree.org/
[3]http://PhDTree.org/

and APS (American Physic Society). In this work, we assume that every advisee has one advisor at the beginning, while multi-tutor situation is not considered. As depicted in Fig. 1, these relationships are hidden in scientific collaboration networks because advisees collaborate with their advisors to publish papers.

Mining advisor-advisee relationship is a challenging problem due to the 'Big Data' nature of modern scholarly data, i.e., the 4 Vs [5]. First (Volume), the increasing volume of scholarly big data makes it difficult to identify such relationships. There are millions of papers, authors, and interaction relationships among them. Second (Variety), existing scholarly data mining methods in relation mining mainly take advantagess of text mining and natural language processing technologies to analyze well-structured data. However, scholarly big data contains heterogeneous information such as authors, papers, venues, and various relationships such as coauthor, citation, and co-work [6]. Third (Velocity), social relationships like advisor-advisee relationships are highly dynamic. One student may have multiple advisors during different time periods such as master versus Ph.D. Meanwhile, an advisee more likely becomes an advisor after graduation. At last (Veracity), the number of co-publications between different advisor-advisee pairs may have a great difference. The lack of ground truth data challenges the traditional supervised learning approaches.

In this work, we propose a novel deep learning model named Shifu ('Shifu' is the Chinese pinyin of 'advisor') to identify every junior scholar's advisor, which employs scholars' local properties as well as the structural features. Shifu is a kind of stacked autoencoder model which considers various factors that determine the advisor-advisee relationships. Stacked autoencoder is one of the typical deep learning models [7], which uses multi-layer architectures to extract inherent features in data. As advisor-advisee relationships are complicated in nature, deep learning architecture can represent input features without prior knowledge. Hence, we take advantages of the stacked autoencoder model to learn and identify advisor-advisee relationships. Additionally, we crawl advisor-advisee pairs from the PhDTree project to train the proposed model. Unlike prior studies, in this paper, we propose a number of novel features, e.g., collaboration duration and academic age, in the deep learning framework to identify social relationships. Through extensive results, we show that Shifu can achieve a precision of 0.94, which is higher than some exsiting machine learning methods.

The main contributions of this work are as follows:

- **Problem Formulation.** We formulated the problem of advisor-advisee relationships identification and calculated scholars' local properties (academic age, number of publications, collaboration times etc.) and advisor-advisee similarity (academic age difference and cohesion).
- **Benchmark Dataset.** We crawled advisor-advisee pairs from Academic Genealogy Wiki project and extracted their publication information by matching them with DBLP dataset as the benchmark dataset.
- **Mining Algorithms.** We presented the first deep-learning based advisor identifying model, called Shifu and conducted extensive experiments to verifying the performance of proposed method.
- **New Knowledge.** We applied Shifu to the whole DBLP dataset to generate a large-scale advisor-advisee dataset for future studying, which is an enrichment of advisor-advisee relationships on the entire DBLP dataset.

The rest of the paper is organized as follows. Related work is discussed in the next section. Section 3 discusses the design of our proposed model. Section 4 discusses the details of experimental results. Finally, section 5 concludes this paper.

2. RELATED WORK

In this section, we discuss our work in light of the related literature on academic collaboration networks, deep learning model, and relation mining respectively.

2.1 Scientific Collaboration Networks

Modern science is becoming more and more collaborative and scholars from different disciplines start collaborating frequently to write papers. Scientific collaboration, which is presented by the coauthor relationship, is a strong social relationship. Advisor-advisee relationship is hidden in scientific collaboration. Based on the coauthorship we can construct a scientific collaboration network [8]. In scientific collaboration networks, two scholars are considered connected if they coauthored a paper. Understanding the structure and social rules of scientific collaboration networks is of great importance and extensive studies had been done on this topic.

Newman [9, 10] analyzed the collaboration networks in Biology, Medicine, Physics, and Computer Science by using author attributions from papers and preprints appearing in these fields over a 5-year period from 1995 to 1999. This work presented the first look at the collaboration networks and discussed many theoretical measures such as clustering, the giant component, centrality, and shortest path. In this paper, we utilize the scientific ego network [11] to mine advisor-advisee relationships. From an ego perspective of life-long scientific collaboration networks, Peterson [12] analyzed the 473 collaboration profiles and 166,000 collaboration records. The research results revealed that scientific collaboration networks are dominated with weak ties characterized by high turnover rates.

2.2 Deep Learning

Deep learning allows computational models which consist multiple processing layers to represent data with multiple levels of abstraction [13]. It has drawn a lot of academic and industrial interests including speech recognition, visual object recognition, and object detection [14, 15]. Deep learning algorithms could discover the inherent structure in large datasets by using back propagation algorithm to optimize the machine learning parameters that are used to define the representation in each layer from the representation in the previous layer.

In this paper, a typical deep learning architecture called stacked autoencoders [16] is employed as the basis of Shifu. We believe that deep learning will succeed in identifying advisor-advisee relationships because it requires very little engineering. It can easily take advantages of increasing computation and data. Meanwhile, the representation ability of deep learning can easily settle the complex and logistical relationships among input features.

2.3 Relationship Identification

Scholars are embedded in academic social networks and therefore are connected with other scholars in different relationships such as colleague and schoolmate. Advisor-advisee relationship identification is within the scope of relation mining [17]. Relation mining is a crucial issue in social network analysis. Relation mining mainly employs the network theory and language processing technique on text data and structured data including user profiles, online social media, and digital libraries.

Given the importance of social relationship identification and analysis, many studies have been done on the identification of intimate relationship [18, 19, 20]. Diehl et al. proposed to identify the formation and evolution of online user relationships to discover collaboration networks [18]. They formulated the relationship identification problem as identifying relevant communications that substantiate a given social relationship type. Oloritun et al. [19] identified close friendship ties via interactions in different periods, spatial proximity, and gender similarity using logistic regression on time-resolved face-to-face interactions. Steurer et al. leveraged [20] online social network data and location-based data to classify relationship as either partners or acquaintances using both supervised and unsupervised methods. The paper presented by Wang et al. [17] is closely related with our work. However, it did not consider the local properties of scholars such as academic age and number of publications. In addition, their method lacks ground truth data for training and evaluation.

3. DESIGN OF SHIFU

The Shifu advisor-advisee relationship mining model is based on the assumption that advisor-advisee relationships can be estimated or inferred by leveraging the coauthor network. Usually, a junior scholar will collaborate with his/her advisor at the beginning of academic career. These collaborations can be located in digital libraries such as DBLP and AMiner. Shifu takes advantage of stacked autoencoder, which is a typical framework of deep learning, as the basic model for learning and training advisor-advisee relationships. Meanwhile, we extract factors that determine an advisor-advisee relationship from the coauthor network as the input of Shifu. Furthermore, in order to train Shifu, we crawl advisor-advisee pairs from universities in the field of Computer Science from Academic Genealogy Wiki project and match these advisor-advisee pairs with DBLP dataset to get advisees' collaboration networks. After learning and training, the Shifu is applied to the whole DBLP dataset to gain all the advisor-advisee pairs in Computer Science and an advisor-advisee dataset can be gained. The idea of designing Shifu is presented in Fig. 2.

3.1 Problem Formulation

In this work, we aim to find advisor-advisee pairs from scholarly big data, i.e., DBLP. In general, every scholar gets coached at the early stage of his/her academic career. For example, in China, PhD students take 3 to 8 years to get graduation. Thus, we are required to focus on the publication information at the early stage to find out their advisors.

We assume there is a scholar i and a set of his/her collaborators $J = \{j_1, j_2...j_n\}$ in DBLP where one of J is the advisor of i. The multi-tutor situation is not considered

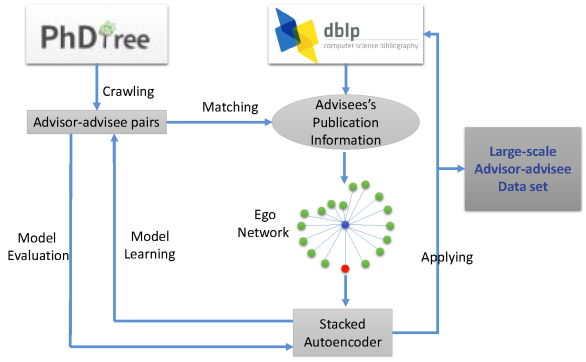


Figure 2: The framework of Shifu.

in this work because few students will have more than one advisors (which is verified in the training data).

We define and calculate the advisor-advisee strength $S_{i,j}$ where each s_{ij} is a set of features $\{s^1, s^2...s^v\}$ that determines an advisor-advisee relationship. For example, s^1 can be the collaboration times between i and j. Then, the task becomes finding a suitable way to present $S_{i,j}$ for every potential advisor-advisee pairs i, j. The advisor-advisee relationship mining problem can be formulated as:

Input: A scholar i, the set of his/her collaborators J, and advisor-advisee strength $S_{i,j}$.

Output: Whether j is i's advisor?

Meanwhile, another important problem we have to solve is to find a collection of advisor-advisee pairs as the ground truth for model training and evaluation.

3.2 Model Description

The Shifu model is built based on the stacked autoencoders, which uses autoencoders as the foundation to create a deep network. There are mainly three steps to build the Shifu: unsupervised feature learning, hidden layer training, and supervised fine tuning.

3.2.1 Unsupervised Feature Learning

Since the advisor-advisee relationship is complex and time-varying, it is difficult to manually label what features will determine such relationship. We want to use autoencoders to represent raw input features. Thus, Shifu uses the autoencoder to reproduce its input to get better feature representation with unsupervised learning. An autoencoder consists of one input layer, one hidden layer, and one output layer. The input layer is a set of training samples $x = \{x^{(1)}, x^{(2)}, x^{(3)}, ...x^{(m)}\}$. The autoencoder aims at finding a suitable function $h_{W,b}(X)$ to make $z = \{z^{(1)}, z^{(2)}, ...z^{(m)}\} \simeq x$. In order to get such function, it has two steps, coding and decoding. The coding step encodes the input x to a hidden representation $y(x^{(i)})$ based on a coding function (Equation 1).

$$y(x) = f(W_1 x + b_1) \tag{1}$$

where W_1 is the coding matrix, b_1 is the coding bias vector. The decoding step decodes the representation $y(x^{(i)})$ back to $x^{(i)}$ with a reconstruction function (Equation 2).

$$z(x) = g(W_2 y(x) + b_2) \tag{2}$$

where W_2 is the decoding matrix, b_2 is the decoding bias vector. Meanwhile, similar with other activation functions in neural networks [7], the activation function of $f(x)$ and $g(x)$ in this paper is logistic sigmoid function:

$$f(x)/g(x) = \frac{1}{1+exp(-x)} \tag{3}$$

The reason that we adopt this function as activation function is that its derivative is:

$$f'(x) = f(x)(1-f(x)) \tag{4}$$

The model parameters can be calculated by minimizing the cost function $L(W,b)$, which can be calculated as:

$$\begin{aligned} L(W,b) &= \frac{1}{m}\sum_{i=1}^{m} L(W,b,x^{(i)},y^{(i)}) \\ &= \frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2} \parallel x^{(i)} - z(x^{(i)}) \parallel^2) \end{aligned} \tag{5}$$

3.2.2 Hidden Layer Training

In the unsupervised feature learning section, we define a hidden layer which can represent the input layer with fewer hidden units. After that, the Shifu is created by stacking autoencoders to form a deep network with more than one hidden layer. As can be seen from Fig. 3, considering Shifu with t hidden layers, the first hidden layer is trained as an autoencoder with the input training set X. After getting the first hidden layer $Y(x)$, the $(k+1)$th hidden layer takes the output of kth hidden layer as the input. In this way, a multi-layer deep network can be stacked hierarchically.

However, if the units of hidden layers are the same or larger than the input layer, autoencoder could potentially learn the identify function [7]. Thus, researchers introduced sparsity constrain into autoencoders to solve this problem [21]. In sparsity autoencoders, let $a_j^{(i)}(x)$ be the activation degree of the units j of hidden layer i, the average activation degree of hidden layer ρ_j can be calculated as:

$$\rho_j' = \frac{1}{m}\sum_{i=1}^{m}[a_j^{(i)}(x^{(i)})] = \rho \tag{6}$$

where ρ is the sparse parameter which is a small value close to 0 (e.g., 0.05). In order to achieve such sparse constrain, we need to introduce another penalty factor to maintain $\rho_j' = \rho$. The penalty factor can be calculated based on the Kullback-Leibler (KL) divergence, which is defined as

$$\sum_{j=1}^{N} KL(\rho_j' \parallel \rho) = \sum_{j=1}^{N}[\rho \log \frac{\rho}{\rho_j'} + (1-\rho)\log\frac{1-\rho}{1-\rho_j'}] \tag{7}$$

where N is the number of units in hidden layer i. Thus, the cost function can be calculated as:

$$L_{saprse}(W,b) = L(W,b) + \beta\sum_{j=1}^{N} KL(\rho_j' \parallel \rho) \tag{8}$$

3.2.3 Supervised Fine Tuning

After previous two steps, we can gain a better representation of the input features. However, all these procedures are unsupervised, which can not be used for advisor-advisee identification. To use these stacked autoencoders for relationship identification, we need to use a supervised classifier on the last hidden layer. As can be seen from Fig. 3, the classifier takes the output of the last hidden layer as input and outputs the classification results. In this paper, we use a logistical regression as the classifier.

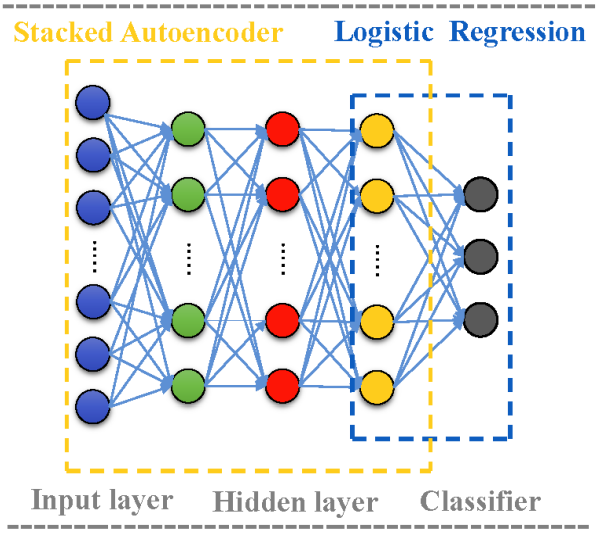


Figure 3: Deep architecture of Shifu.

3.3 Model Learning

In order to minimize the cost function, the BP (Back Propagation) method with gradient-based optimization method is employed. Specifically, we adopt the greedy layerwise unsupervised learning algorithm proposed by Hinton et al [22]. The main idea is to train the deep networks layer by layer in a bottom-up order. The procedure is based on the idea in [7] which can be described as follows.

- Train the parameters of the first hidden layer by minimizing the cost function where the training data is the input data.
- Train the second hidden layer by taking the output of first hidden layer as the input.
- Repeat the second step for all the hidden layers.
- Train the classifier by taking the output of last hidden layer as the input in a supervised way.
- Fine-tune the parameters with the BP method in a supervised way.

3.4 Input Features

In order to apply the Shifu model to mine advisor-advisee relationships, we need to input the collaboration information between advisees and their collaborators. For the input features, we consider all the possible factors determining an advisor-advisee relationship which can be extracted from DBLP dataset. We consider the local properties of advisees and their collaborators, and similarity between them. Specifically, given an advisee i and his/her collaborators $J = \{j_1, j_2 ... j_n\}$, we mainly consider the local features and network features as shown in Table 1.

The academic age of a given scholar is calculated by the investigated year minus the year when he/she published first

Table 1: Descriptions of input features

Feature	Description
AA_i	academic age of i when first collaborating with j
N_i	No. of i's publication before collaborating with j
AA_j	academic age of j when first collaborating with i
N_j	No. of j's publication before collaborating with i
AD	academic age difference value between i and j
CT	collaborating times between i and j
CD	collaborating duration between i and j
FTA	number of times i and j being first two authors
$Cohesion$	similarity between i and j (first 8 years)

paper. Based on the assumption that an advisee is coached by his/her advisor at the beginning of the academic career, we calculate all these input features based on publication information during the first eight years. Since the cohesion between two collaborators is time-varying, there are eight different cohesion values during the investigated eight years. Meanwhile, in order to improve the learning efficiency, we normalize all the input features into $[0, 1]$ with the min-max normalization approach.

4. EXPERIMENTAL RESULTS

In this section, we present extensive experimental results of evaluating the performance of Shifu. We describe how to get the suitable experimental dataset from PhDTree project and DBLP. We compare the proposed Shifu model with various machine learning methods including Logistic Regression (LR), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), and TPFG [17].

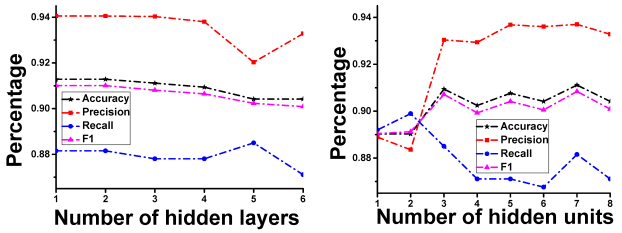
4.1 datasets

We use the PhDTree project, which is a crowd-sourced wiki website to document the academic family tree of PhDs worldwide, to gain the ground truth advisor-advisee pairs. We crawl the advisor-advisee pairs in the field of Computer Science. Since PhDTree contains little publication information of a given advisee beside advisor-advisee relationships, we match the advisee to the DBLP Computer Science Bibliography Database to further gain their collaboration records as well as publication information. Specifically, we crawl the advisor-advisee pairs from sixteen famous universities such as Carnegie Mellon University and Stanford University.

For each advisor-advisee pair, we first search and identify advisors in PhDTree project and the advisees of the selected advisors are crawled from their CVs. When we match the name of an advisee, we use the regular expression to present the name. For each advisee, we further match their information in the DBLP dataset both considering the name of the advisee and the name of his/her advisor. In other words, only if we can match both the name of a given advisee and his/her advisor, we define these two collaborators as the ground truth advisor-advisee relationship pairs. Thus, the name ambiguity problem can be solved. After name matching, we crawled 3,423 advisor-advisee relationships. For each advisee, we calculate the input features based on the publication information in DBLP.

4.2 Parameter Settings

Considering the structure of Shifu, we need to determine the number of hidden layers L and the number of hidden units in each hidden layer U. Parameter setting is an important procedure since Shifu has good self-learning ability which heavily relies on the number of hidden layers and the number of hidden units. Although neural networks with single hidden layer may have good learning ability, more hidden layers will have better performance. However, as the number of hidden layers and units increases, the number of parameters will increase accordingly. If there are too many hidden layers, it is difficult to use the BP method to train the model because the error will diverge. In this paper, we choose L from 1 to 6 and the number of U from 1 to 8. After experiments, as can be seen from Fig. 4, we obtain the best architecture which consists of three hidden layers, and the number of units in each hidden layer is $[7, 7, 7]$. For instance, such parameter setting has been proofed effectively in [7].



(a) Hidden Layers (b) Hidden Units

Figure 4: Performance of Shifu in terms of different hidden units and hidden layers.

4.3 Results

Since the advisor-advisee identification is a dichotomy problem, i.e., j is the advisor of advisee i or j is not the advisor of i, we choose four popular metrics, accuracy, precision, recall, and F1 to evaluate the performance of Shifu. We randomly select 90% advisor-advisee pairs in the datasets as the training dataset and the rest as the testing dataset. Meanwhile, in order to test the stability and fidelity of Shifu, we use k-fold (k=10)cross validation in each experiment.

4.3.1 *Effect of Time Length*

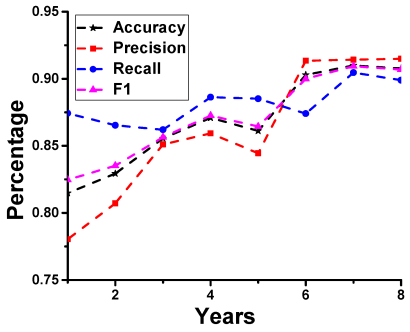


Figure 5: Performance of Shifu over different time length of input data.

Advisees may encounter with their advisors at different academic ages. One the one hand, advisees may collaborate with their advisors at the very beginning. On the other hand, advisees may start collaborating with their advisors after publishing several papers. It is difficult to find out

when an advisee will collaborate with his/her advisor or how long they will collaborate. To examine the influence of different data time length, we conduct experiments in this part and the time length ranges from 1 to 8.

As shown in Fig. 5, if we merely consider the first academic age of advisees, the accuracy, precision, recall, and F1 are very low. The overall trend of four evaluation metrics increases with more investigated years. It is noticeable that if we use the advisees' publication information of first 6 years, the Shifu can achieve relative better performance. The reason accounting for this phenomena is that on average it takes about 5 years for a PhD student to get graduated. Meanwhile, there is a time delay for publication. For example, an advisee may write one paper during PhD project and this paper may get published after graduation. We thus take advantages of the first-eight-year publication information for each advisee to achieve better results in this paper.

4.3.2 *Effect of Academic Age*

Previous research on scientific collaboration network analysis mainly focuses on network properties from the macroscopic perspective [23, 24]. However, from the scholarly big data, we can infer scholars' local properties and features from the microscopic perspective. We believe that exploiting this local information and properties can benefit the relationship identification. In Shifu model, we consider the academic age, which is an important human demographic property to measure the similarity between advisees and advisors. Since the TPFG method does not take the academic age into consideration, we do not adopt the TPFG method to evaluate the effect of academic age.

The effect of academic age on the performance of Shifu in comparison with the other algorithms can be seen in Fig. 4.3.1. We can see from Fig. 6(a), 6(b), and 6(c) that the precision, recall, and F1 rate are 94.1%, 92.3%, and 92.8% respectively, which means that Shifu can effectively identify the advisor-advisee relationships. Meanwhile, Shifu performs better than KNN, SVM, and LR in precision, recall, and F1 both with or without the factor of academic age. Take precision rate as an example, Shifu is 2.4%, 3.9%, and 4.7% higher than KNN, SVM, and LR respectively. Meanwhile, experimental results show that academic age is an important feature for all the experiments and baseline methods. Another conclusion we can gain from this figure is that, these machine learning algorithms can accurately ($> 90.1\%$ in precision) identify the advisor-advisee relationships. This observation confirms the assumption that advisor-advisee is a strong social relationship which is hidden in scientific collaboration networks.

4.3.3 *Effect of Training Data Size*

The ground truth dataset contains 3,423 advisor-advisee pairs. In the experiments, we divide the dataset into two subsets, training set and testing set. In order to explore the performance of Shifu regarding to the size of training dataset, we use different proportions of training dataset.

As shown in Fig. 7, Shifu performs better than other machine learning algorithms in terms of precision, recall, and F1. Take the 90% fraction of training set as example, we can see that Shifu is 3.5%, 4.2%, 4.4%, and 3.0% higher in precision than KNN, SVM, LR, and TPFG respectively. Another observation from Figs. 7(a), 7(b), and 7(c) is that as the fraction of training dataset increases, all methods tend to have better performance. In our experiment, although there are 3,423 advisor-advisors pairs, it would be better if we take advantages of more ground truth datasets.

4.3.4 *Effect of Input Features*

Fig. 8 depicts the precision distribution regarding six important input features and the node distribution of each input features. The size of the dot in each subfigure represents the number of scholars at this point. Bigger circles indicate that there are more fractions of nodes, whereas smaller circles mean less fractions of nodes.

Fig. 8(a) shows the distribution of advisor's AA and the precision of advisor-advisee relationship identification at different AA. It is clear that most advisors' academic ages range from 10 to 20 and fewer professors will continue directing PhD students after academic age 35. The Shifu has higher precision if the AA of the advisor is more than 10. If the academic age of an advisor is samller than 5, it is difficult for Shifu to identify his/her advisor-advisee relationship. The good news is that there are few such young advisors. Meanwhile, if the advisor's AA is higher than 30, Shifu can reach a precision of 100%. Fig. 8(b) depicts the effect of different academic ages on precision of Shifu in identifying advisee-advisors relationships. As illustrated in this figure, the trend of precision goes up with higher academic age differences, which is similar with that of AA. Fig. 8(c) depicts the precision distribution in terms of advisors' publication numbers. As can be seen in this figure, most advisors have about fifty publications and Shifu performs better in identifying advisors with more publications. We can also notice from this figure that several advisors have only one publication. In this case, Shifu can hardly identify these advisor-advisee pairs. The effect of other input features is shown in the rest subfigures.

4.4 Applications

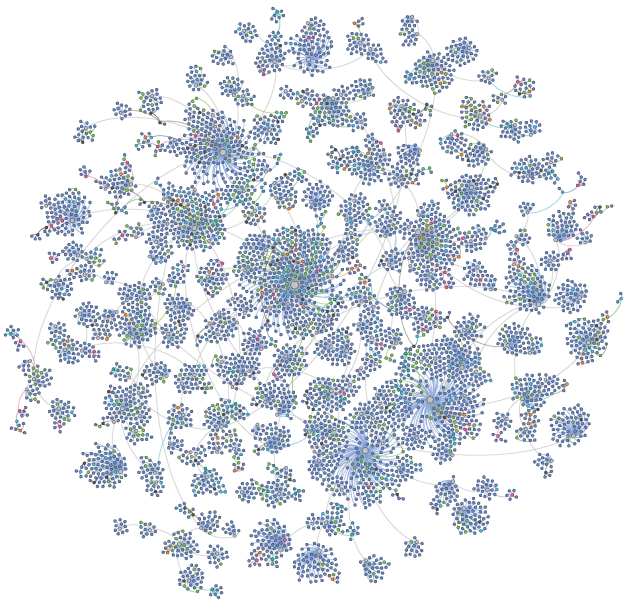
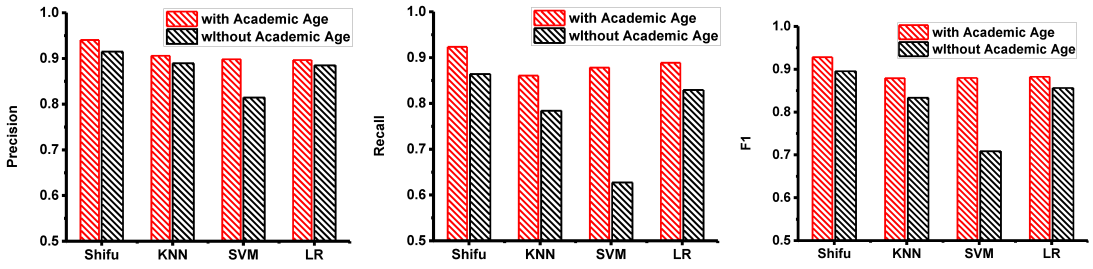
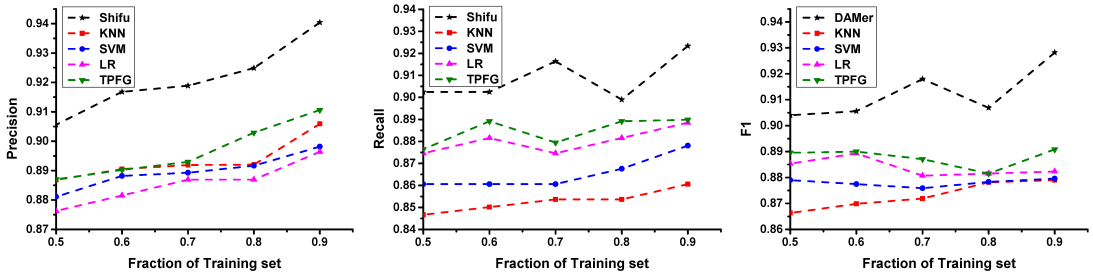


Figure 9: Visualization of the largest connected component in the DBLP genealogy.



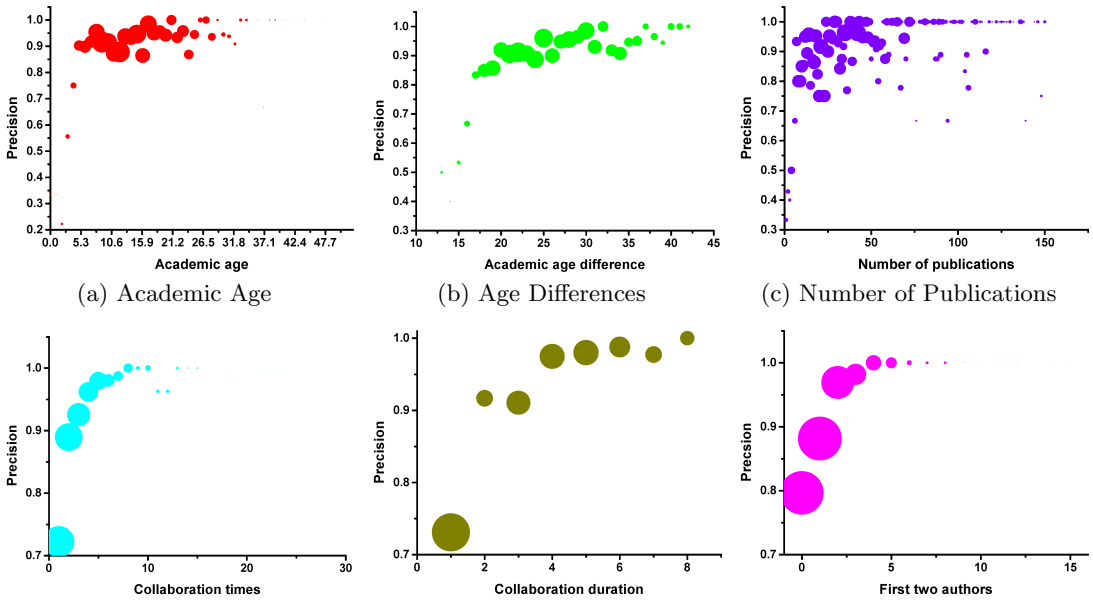
(a) Precision (b) Recall (c) F1

Figure 6: Comparison of Shifu with/without the factor of academic age in terms of precision, recall, and F1.



(a) Precision (b) Recall (c) F1

Figure 7: Comparison of Shifu with other methods in terms of precision, recall, and F1.



(a) Academic Age (b) Age Differences (c) Number of Publications

(d) Collaboration Times (e) Collaboration Duration (f) First Two Author Times

Figure 8: The effect of input features on the performance of Shifu in terms of precision.

Finally, we apply Shifu to the whole DBLP dataset. For a given scholar in DBLP, we calculate the input features based his/her first-eight-year publication information. Then we run Shifu for each scholar. Thus, we can gain a large-scale advisor-advise pair dataset, which is an enrichment of advisor-advisee relationships on the entire DBLP dataset. From preliminary results, we have inferred 1, 111, 513 advisor-advisee pairs. For example, Fig 9 shows a subtree of DBLP genealogy, where the color represents a family of scholars. We can see the advisor-advisee relationships in a chronological hierarchy. This advisor-advisee can also be used for many theoretical and practical applications, such as visualization of academic genealogy, mentor performance evaluation, and anatomy of advisor-advisee relationships.

5. CONCLUSION

We have studied the problem of mining the advisor-advisee relationships from scholarly big data as attempt to automatically generate a large-scale advisor-advisee dataset. Based on the assumption that advisor-advisee relationships are hidden in coauthor networks, we propose a novel method named Shifu based on deep learning algorithms. We define and calculate a number of novel features to measure the similarity between advisors and advisees extracted from scientific collaboration networks. In order to train and evaluate the proposed model, we crawl advisor-advisee pairs from PhDTree project and match them with DBLP digital libraries to gain the ground truth dataset. Finally, extensive experimental results demonstrate the effectiveness of our proposed model. Future work includes generalizing Shifu to other disciplines.

6. REFERENCES

[1] R Dean Malmgren, Julio M Ottino, and Luís A Nunes Amaral. The role of mentorship in protégé performance. *Nature*, 465(7298):622–626, 2010.

[2] Georgia T Chao, Patm Walz, and Philip D Gardner. Formal and informal mentorships: A comparison on mentoring functions and contrast with nonmentored counterparts. *Personnel Psychology*, 45(3):619–636, 1992.

[3] Tammy D Allen, Mark L Poteet, Joyce EA Russell, and Gregory H Dobbins. A field study of factors related to supervisors' willingness to mentor others. *Journal of Vocational Behavior*, 50(1):1–22, 1997.

[4] Stewart I Donaldson, Ellen A Ensher, and Elisa J Grant-Vallone. Longitudinal examination of mentoring relationships on organizational commitment and citizenship behavior. *Journal of Career Development*, 26(4):233–249, 2000.

[5] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pages 42–47. IEEE, 2013.

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

[7] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2015.

[8] M. E. J Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205, 2004.

[9] Mark Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[10] M. E. J Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132, 2001.

[11] Mark Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25(1):83–95, 2003.

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

[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[14] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[15] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[16] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine learning*, pages 1096–1103. ACM, 2008.

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

[18] Christopher P Diehl, Galileo Namata, and Lise Getoor. Relationship identification for social network discovery. In *AAAI*, volume 22, pages 546–552, 2007.

[19] Rahman O Oloritun, Anmol Madan, Alex Pentland, and Inas Khayal. Identifying close friendships in a sensed social network. *Procedia-Social and Behavioral Sciences*, 79:18–26, 2013.

[20] Michael Steurer and Christoph Trattner. Acquaintance or partner?: predicting partnership in online and location-based social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 372–379. ACM, 2013.

[21] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010.

[22] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[23] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 1285–1293. ACM, 2012.

[24] Giseli Rabello Lopes, Mirella M Moro, Leandro Krug Wives, and José Palazzo Moreira De Oliveira. Collaboration recommendation on academic social networks. In *Advances in Conceptual Modeling–Applications and Challenges*, pages 190–199. Springer, 2010.