Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 8135464, 9 pages
https://doi.org/10.1155/2017/8135464

Research Article

Inferring Advisor-Student Relationships from Publication Networks Based on Approximate MaxConfidence Measure

Yongjun Li,[1] Nan Fang,[2] Zun Liu,[1] and Hui Yu[1]

[1] *School of Computer Science & Technology, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China*
[2] *CNOOC Information Technology Limited, Beijing 100010, China*

Correspondence should be addressed to Yongjun Li; lyj@nwpu.edu.cn

Received 29 September 2016; Revised 9 February 2017; Accepted 16 February 2017; Published 6 March 2017

Academic Editor: Erik Cuevas

A publication network contains abundant knowledge about advisor-student relationships. However, these relationship labels are not explicitly shown and need to be identified based on the hidden knowledge. The exploration of such relationships can benefit many interesting applications such as expert finding and research community analysis and has already drawn many scholars' attention. In this paper, based on the common knowledge that a student usually coauthors his papers with his advisor, we propose an approximate *MaxConfidence* measure and present an advisor-student relationship identification algorithm based on the proposed measure. Based on the comparison of two authors' publication list, we first employ the proposed measure to determine the time interval that a potential advising relationship lasts and then infer the likelihood of this potential advising relationship. Our algorithm suggests an advisor for each student based on the inference results. The experiment results show that our algorithm can infer advisor-student relationships efficiently and achieve a better accuracy than the time-constrained probabilistic factor graph (TPFG) model without any supervised information. Also, we apply some reasonable restrictions on the dataset to reduce the search space significantly.

1. Introduction

Online social networks, such as Facebook and Twitter, have become popular in our daily lives. By the social network, we can establish and maintain social relationships with others or share information with them. It is generally accepted that different relationships have essentially different influence on people. For example, the advisor largely influences a Ph.D. candidate's research field, while his classmates or families largely influence his hobbies. Unfortunately, in online social networks, these relationship labels are often hidden and only a few users label their relationships. Statistics show that less than 23% relationships on LinkedIn have been labeled [1]. Recently, the relationship identification has drawn many scholars' attention. Awareness of these relationship types can benefit many applications. For example, the advisor-student relationship is helpful to discover conflict of interests in the review process for research papers or projects.

There are several works that focus on relationship identification. Diehl et al. [2] try to identify the manager-subordinate relationships by learning a ranking function. Eagle et al. [3] present several patterns discovered in mobile phone data and try to infer the friendship network. However, these works focus on special domains which are different from our work. Wang et al. [4] propose a deep learning based advisor-advisee relationship identification method which considers the personal properties and network characteristics with a stacked autoencoder model. Wang et al. [5] employ the factor graph and propose an unsupervised probabilistic model, named TPFG, for mining the advisor-advisee relationships from the publication network. Tang et al. [6] develop a framework for classifying the type of social relationships by learning across heterogeneous networks. Zhuang et al. [7] precisely define the problem of inferring social ties and propose a Partially Labeled Pairwise Factor Graph Model (PLP-FGM) for learning to infer the type of social relationships. Tang et al. [8, 9] and He et al. [10] present a framework for classifying the type of social relationships by learning across heterogeneous networks, respectively. These proposed algorithms are based on factor graph and are computation-intensive.

In this paper, we aim to propose an easily calculated and effective algorithm to solve this problem. Based on the common knowledge that a graduate student always coauthors his papers with his advisor but not vice versa, we employ an approximate *MaxConfidence* measure to determine the time interval that a potential advising relationship lasts and then infer the probability of this potential advising relationship. Our work does not need any supervised information. Experiment results show that our approach can achieve a better accuracy than TPFG [5] does.

The rest of this paper is organized as follows. Section 2 formally formulates the problem, and the proposed approach is detailed in Section 3. The experiment results are presented to validate the efficiency and effectiveness of our methodology in Section 4. We present the related works in the Section 5 and conclude this paper in Section 6 with remarks and future works.

2. Problem Formulation and Restrictions

A publication network is represented as a bipartite graph $G = (V^p, V^a, E)$, where V^p is the set of publications, $V^a = \{a_1, a_2, \ldots, a_n\}$ is the set of authors, and E is the set of edges. Let vector $L_i = \{l_{i1}, l_{i2}, \ldots, l_{ik}\}$ denote a_i's publication list, where each $l_{ik} \in V^p$. n_i^y indicates the number of papers a_i has published by the year y, and n_{ij}^y indicates the number of papers a_i and a_j have coauthored by the year y.

The identification result of our method is represented by "*advisor-student relationship,*" which is defined as follows.

Definition 1 (advisor-student relationship). *Advisor-student relationship* is a triple tuple (p_{ij}, s_{ij}, t_{ij}), where p_{ij} is the relationship score of the author a_i advising the author a_j and s_{ij} and t_{ij} are years the relation starts from and terminates at.

Based on the above definitions, we can formulate the problem of our work. Given a publication network G, our research objective is to uncover all the advisor-student relationship hidden in G.

Problem 2 (advisor-student relationship identification). Given a publication network G, the objective is to find a relationship identification algorithm

$$f : G \longrightarrow R, \tag{1}$$

where R is an advisor-student relationship matrix. Each element of this matrix is triplet showing the advisor-student relationship (p_{ij}, s_{ij}, t_{ij}). In the proposed approach, our main focuses are on how to infer p_{ij}, s_{ij}, and t_{ij}, respectively.

To reduce the search space and the calculation time, we first make some restrictions to simplify our work. In the following restrictions, we assume a_i is the advisor candidate of a_j.

Restriction R1 ($\forall a_i \in V^a$, $|L_i| \geq 3$). We assume that each author, including advisor and student, should publish at least three papers. If an author only publishes one or two papers, it is very difficult to identify his advisor even if we do it manually. We check each author's publication list and only keep the authors who do not violate R1. In the experiment section, we test this restriction and show its effect.

Restriction R2 ($\forall a_i, a_j \in V^a$, $[s_i, t_i] \cap [s_j, t_j] \neq \emptyset$). $[s_i, t_i]$ is the interval representing the publication history of author a_i. Exactly, this restriction is a natural consequence of time causality. It reflects the following fact that a student's publication history should have an intersection with his potential advisor's publication history. We will use this restriction to filter out those unlikely advisor-student relations.

Restriction R3 ($t_{ij} - s_{ij} > 1$). We also restrict that a student spends more than one year on his degree. If the advisor-student relationship only lasts for no longer than one year, it is impossible to distinguish the advisor-student relationship from temporary partnership.

3. *MaxConfidence*-Based Approach

We first propose an approximate *MaxConfidence* measure and present an advisor-student relationship identification algorithm based on the *MaxConfidence* measure. Given author a_i, a_j and their respective publication list L_i, L_j, we present how to infer s_{ij}, t_{ij}, and p_{ij}, respectively. Before we compare L_i and L_j, restriction R2 is checked. If and only if R2 is satisfied, a potential advisor-student relationship may exist between a_i and a_j. If $s_j > s_i$, a_i is a potential advisor of a_j, and if $s_i > s_j$, a_j is a potential advisor. In the following paragraph, we assume a_i is a_j's potential advisor.

The starting year of the advisor-student relationship, s_{ij}, is estimated as the year when a_i and a_j coauthored their first paper. In general, the estimated start year lags behind the actual time, because the advisor and his students usually coauthor their first paper after the advisor-student relationship has been established.

Before describing the estimation method of t_{ij}, we define an approximate *MaxConfidence* measure for the coauthored publications of a_i and a_j. The original version of *MaxConfidence* measure can be found in [11], as shown in

$$\begin{aligned} & MaxConfidence\,(m,k) \\ & \quad = \max\left\{\frac{\sup(mk)}{\sup(m)}, \frac{\sup(mk)}{\sup(k)}\right\}. \end{aligned} \tag{2}$$

Given two arbitrary authors m and k, $\sup(m)$ or $\sup(k)$ denote the number of papers published by m or k. $\sup(mk)$ denote the number of papers coauthored by m and k. *MaxConfidence* $(m, k) = 1$ reflects a slightly weaker association condition where one author always coauthors with the other but the converse may not necessarily be true. In an advisor-student relationship, the student always coauthors with his advisor, but conversely, the advisor only coauthors a small portion of his papers with this student [11]. Thus, if a_i is a_j's advisor, *MaxConfidence* (a_i, a_j) should be close to 1; while a_i and a_j are only collaborators, *MaxConfidence* (a_i, a_j) may not be close to 1. So, the *MaxConfidence* measure reflects the high correlation between the advisor's publication and his students'. However, *MaxConfidence* measure is symmetrical;

that is, *MaxConfidence* (a_i, a_j) = *MaxConfidence* (a_j, a_i). We cannot distinguish who is an advisor and who is a student based only on *MaxConfidence* measure. For clear distinction between advisor and student, we define an approximate *MaxConfidence (AMC)* measure to describe the correlation between the advisor's publication and his students'. The proposed measure is defined as shown in

$$H_{ij}(t) = \begin{cases} \dfrac{n_{ij}^{y}}{n_j^{y}}, & \text{if } n_j^{y} \neq 0, \\ 1, & \text{if } n_j^{y} = 0 \wedge t \leq t_j, \\ 0, & \text{if } n_j^{y} = 0 \wedge t > t_j. \end{cases} \quad (3)$$

Here we introduce the time factor into *AMC* measure to calculate it year by year and determine the end year of the advisor-student relationship according to the change of measure $\{H_{ij}\}_t$.

According to definition, *AMC* measure is asymmetrical. In general, student a_j usually coauthors most of all his papers with his advisor a_i, so $H_{ij}(t)$ often approaches 1. Conversely, advisor a_i coauthors only a few of his papers with student a_j, so $H_{ji}(t)$ does not necessarily approaches 1. Based on the property of *AMC* measure described above, the end year of the advisor-student relationship, t_{ij}, can be estimated as the year before the two years in a row with *AMC* measures much less than 1. For simplicity, we use $H_{ij}(t) < \theta$ to express that *AMC* measure becomes much less than 1, where θ is a threshold.

After s_{ij} and t_{ij} are determined, we check them to see if they meet restriction R3. Only if R3 holds is the relationship between a_i and a_j considered as advisor-student relationship. The student a_j may have more than one advisor candidate. In order to easily distinguish from all candidates who really is the advisor of a_j, we calculate a relationship score for every potential advisor-student relationship based on (4). This score reflects the likelihood that a_i advised a_j from s_{ij} to t_{ij}.

$$p_{ij} = 1 - 2e^{n_j - 2n_{ij}}, \quad (4)$$

where n_j indicates the number of papers a_j has published from s_{ij} to t_{ij} and n_{ij} indicates the number of papers a_i and a_j have coauthored from s_{ij} to t_{ij}.

After calculating the score of every potential advisor-student relationship, we select the candidate with the maximum relationship score as the advisor of a_j.

We refer to our algorithm as AMC in the following paragraph. Here AMC is summarized as shown in Algorithm 1.

In AMC, the restriction R2 is checked on every pair of authors. Only those pairs of authors that pass R2 are considered. In the inference, we compare every two authors' publication lists, so the time complexity of AMC is decided by the average number of papers published by each author and the number of author pairs that pass R2. If we assume the average number of papers published by each author is N and the number of author pairs that meet R2 is M, the time complexity is $O(NM)$. M is many orders of magnitude smaller than $|V^a| \times |V^a|$. For example, in one of our datasets, M is 367,543 while $|V^a|$ is 243,537.

As described in [5], TPFG is a two-stage framework, including preprocessing and TPFG model. The purpose of preprocessing is to generate the candidate graph and reduce the search space. The total complexity of this stage is $O(NM)$. In the 2nd stage, a time-constrained probabilistic factor graph model is leveraged to decompose the joint probability of the unknown advisor of every author. By maximizing the joint probability of the factor graph, the authors can infer the advisor-student relationship. Compared with TPFG, the time complexity of AMC is as same as the preprocessing of TPFG. AMC does not perform any further learning after the above processes. However, TPFG has to perform a factor graph based model learning to infer the probability of advising relationship. This learning process is computation-intensive, especially on large scale cycle-containing publication network. Based on the above analysis, we easily figure out that the time complexity of AMC is much lower than that of TPFG.

4. Experiment Results

In our experiment, there are two types of datasets: unlabeled dataset and labeled dataset. The unlabeled dataset contain two datasets, sampled from Digital Bibliography and Library Project (http://dblp.uni-trier.de/) (DBLP) and used to infer the advisor-student relationship. One unlabeled dataset with the time span from 1990 to 2011 is referred to as Datset2011, and the other unlabeled one also used in [5] is referred to as Dataset2008. The basic publication information of each paper is listed in DBLP, which includes all authors, publication year, and publication venue. The labeled datasets are used to test the accuracy of the discovered advisor-student relationships, and these labeled datasets are detailed in experiment setup subsection.

4.1. Effect of Restriction R1. We first take Dataset2011, for example, to test the effectiveness of restriction R1. The complementary cumulative distribution of the number of every author's publications is shown in Figure 1.

From Figure 1, we can see that more than 51% authors only publish one paper and the number of 68% authors' publications is less than 3. That is, after we apply the restriction R1 on the dataset, the new search space will be reduced to 10.24% (32% × 32% = 10.24%) of the original search space. This reduction will significantly save the computation time.

4.2. Experiment Setup. Before feeding Dataset2011 and Dataset2008 to our approach, we employ the restriction R1 to filter them. After filtration, Dataset2011 consists of 996,427 authors and 1,656,588 publications, and Dataset2008 contains 142,717 authors and 969,286 publications. On the other side, the labeled datasets consist of three validation datasets. All of these three datasets come from [5]. As described in [5], one original dataset is manually labeled by looking up advisors' homepage, and the other two are crawled from two websites, Mathematics Genealogy Project and AI Genealogy Project. These datasets are referred to as MAN, MathGP, and AIGP, respectively. The MAN is further divided into three subdatasets: Teacher, PhD, and Colleague. Teacher contains both graduated students and graduate students pairing with

```
Input: $G' = (V^w, V^a, E)$
Output: $R = \{(p_{ij}, s_{ij}, t_{ij})_{i,j \in V^a}\}$
Initialize all $p_{ij}$ as 0, $s_{ij}$ as 0, $t_{ij}$ as 0, respectively;
For each $a_i \in V^a$ do
    update $s_i$ and $t_i$ according to $a_i$'s publication list;
    for each $a_j \in V^a$ do
        update $s_j$ and $t_j$ according to $a_j$'s publication list;
        check the restriction R2 according to $a_i$'s and $a_j$'s publication history;
        if restriction R2 is valid then
            decide who the potential advisor is according to $s_i$ and $t_j$;
            determinte the start year $s_{ij}$ based on $a_i$'s and $a_j$'s publication list;
            calculate all $n_i^y$, $n_j^y$ and $n_{ij}^y$ year by year;
            determinte the end year $t_{ij}$;
            calculate the relationship score $p_{ij}$ of advisor-student relationship;
    select $a_k$ whose $p_{ki}$ is maximum as $a_i$'s advisor;
```

Algorithm 1

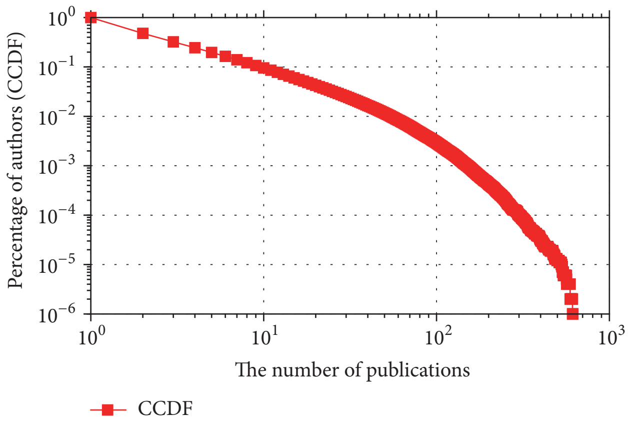


Figure 1: CCDF of the number of author's publications.

their advisors, while PhD only contains advisor-PhD pairs. Colleague is a negative dataset for our experiments, which contains coauthor or colleague pairs. To keep consistency with Data2011 and Data2008, we only select the pairs whose student graduated after 1990 as our labeled data. We also exclude the advisor-student pairs whose advisor is not labeled clearly in the original labeled datasets.

To quantitatively compare our methods with others, we employ True Positive Rate (TPR) and False Positive Rate (FPR) to evaluate our experiment results. For convenience, we write True Positive as TP, False Positive as FP, False Negative as FN, and True Negative as TN. TPR and FPR are defined as shown in (5). Here TP, FN, FP, and TN are the number of True Positive results, False Negative results, False Positive results, and True Negative results, respectively.

$$\text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \qquad \text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})}. \tag{5}$$

For negative samples, we define the accuracy as (1 − FPR).

In our inference results, all the potential advisor-student pairs will be labeled as one of three labels, Yes, No, or Unknown. The label Yes indicates an advisor-student relationship, label No indicates non-advisor-student relationship, and label Unknown means the algorithm is not certain about the type of the target relationship. To evaluate the coverage of AMC, we defined Coverage Rate (CR) as the ratio of the number of Yes and No to the number of total pairs, as shown in (6). Also we define UnCoverage Rate (UCR) as (1 − CR).

$$\text{CR} = \frac{(\text{Yes} + \text{No})}{(\text{Yes} + \text{No} + \text{Unknown})}. \tag{6}$$

4.3. Comparison with TPFG. As shown in [5], TPFG is more efficient and accurate than other methods, such as SVM, Independent Maxima, so we only compare AMC with TPFG to explore the capability of AMC in mining advisor-student relationships. The results of comparisons on both positive samples and negative samples are shown in Figure 2.

From Figure 2(a), we can see that AMC works a little better than TPFG on three positive datasets, PhD, Teacher, and MathGP. On AIGP these two methods have very similar TPR value. When we identify the advisor-student relationships, the main information which we can use is these collaborations between authors. Whether a_i is the advisor of a_j or not is dependent strongly on collaborations of a_i and a_j. The collaborations of a_i and his other collaborators have little relation with the identification result, and so do the collaborations of a_j and his other collaborators. In TPFG, to compute the marginal maximal probability of a_i and a_j, it needs the messages passed from a_j's neighbors. That is, whether a_i is the advisor of a_j or not is dependent not only on collaborations of a_i and a_j but also on collaborations of other authors. We presume that some errors are also passed in message propagation, which leads to the fall in inference accuracy. Another reason is that TPFG labels every pair of advisor-student candidate as Yes or No, while AMC does not. We can easily see that from Figure 2(b); that is, UCRs of TPFG on three labeled datasets are 0. However, all of UCRs of AMC on four labeled

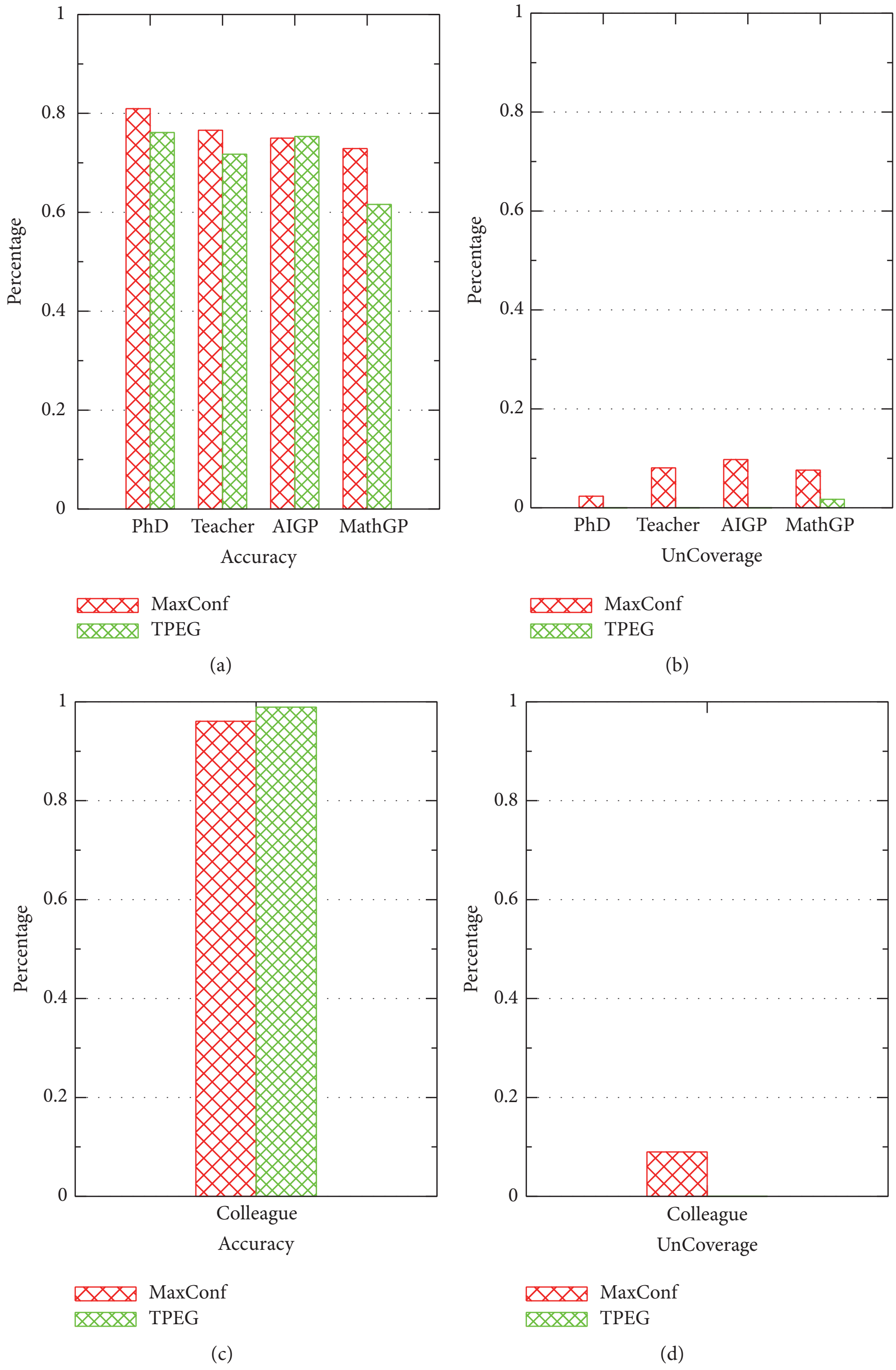


Figure 2: Identification results compared with TPFG. (a) Accuracy on positive samples. (b) UnCoverage on positive samples. (c) Accuracy on negative samples. (d) UnCoverage on negative samples.

datasets are less than 0.1. In other words, AMC identifies more than 90% of the potential advisor-student pairs.

From Figure 2(a) we also see that all TPRs are less than 0.85. The further study on the inference results shows that the error can be contributed to the following reasons. Firstly, some distinct authors may have the same name, especially for Chinese authors. For example, if we try to find the advisor of Chao Liu from UIUC, both TPFG and AMC return wrong results. We further look up the publication history of Chao Liu in DBLP and in his homepage, respectively. Only 2 out of 7 papers are found on Chao Liu's homepage and therefore we can conclude that the remaining 5 papers are from another person who has the same name as Chao Liu from UIUC. Based on his publication history in homepage, Chao Liu's advisor inferred by AMC is Jiawei Han from UIUC. This result is consistent with the labeled dataset. Another major reason can be found from the common phenomenon in DBLP that an author may publish papers using different names. Yan Lindsay Sun and Yan Sun are considered as two distinct authors in our datasets. Based on AMC, we can infer that the advisor of Yan Sun is K. J. Ray Liu, but we cannot find who the advisor of Yan Lindsay Sun is. In fact, Yan Lindsay Sun and Yan Sun are two distinct names of an identical author. Her advisor is K. J. Ray Liu. After graduating from University of Maryland in 2004, she starts to use "Yan Lindsay Sun" as her another name. The above two reasons are called name ambiguity [12]. As described in [5], at least 40% of error is contributed to name ambiguity. The accuracy of all methods could be further improved if the name ambiguity problem was solved perfectly. The third reason is that some students coauthored fewer papers with his advisor and sometimes even have no publication coauthored with his advisor. In these situations, it is almost impossible to identify these latent advisor-student relationships merely based on their publication history.

From Figures 2(c) and 2(d) we find that TPFG works a little better than AMC on negative samples. However, the TPR of AMC is also more than 96% as shown in Figure 2(c). The reason why TPFG works better on negative samples is the same as why it works worse on positive samples.

4.4. Identification Results on Different Datasets. To test whether the inference accuracy will decrease with increase in the number of name ambiguities and whether the possible data sampling bias exists, we apply AMC on Data2008 and Data2011, respectively. The identification results are shown in Figure 3.

From Figures 3(a) and 3(b) we find that the proposed method has higher TPR and lower UCR on Data2008 than on Data2011. We conduct further study on these two datasets to explore the reason why the proposed method works badly on the big dataset. The number of authors of Data2011 is nearly seven times that of Data2008. The probability of name ambiguity increases as the number of authors increases. This leads to the reduction in inference accuracy. In negative samples, our algorithm works a little better on Data2011 than on Data2008, but no significant changes were shown from Figures 3(c) and 3(d).

4.5. Estimated End Time. In this subsection, we show the performance of the inference on t_{ij} by AMC. Figure 4 shows the distribution of deviation of t_{ij}. The average deviation of t_{ij} is 1.39. The median of these deviations is 1.

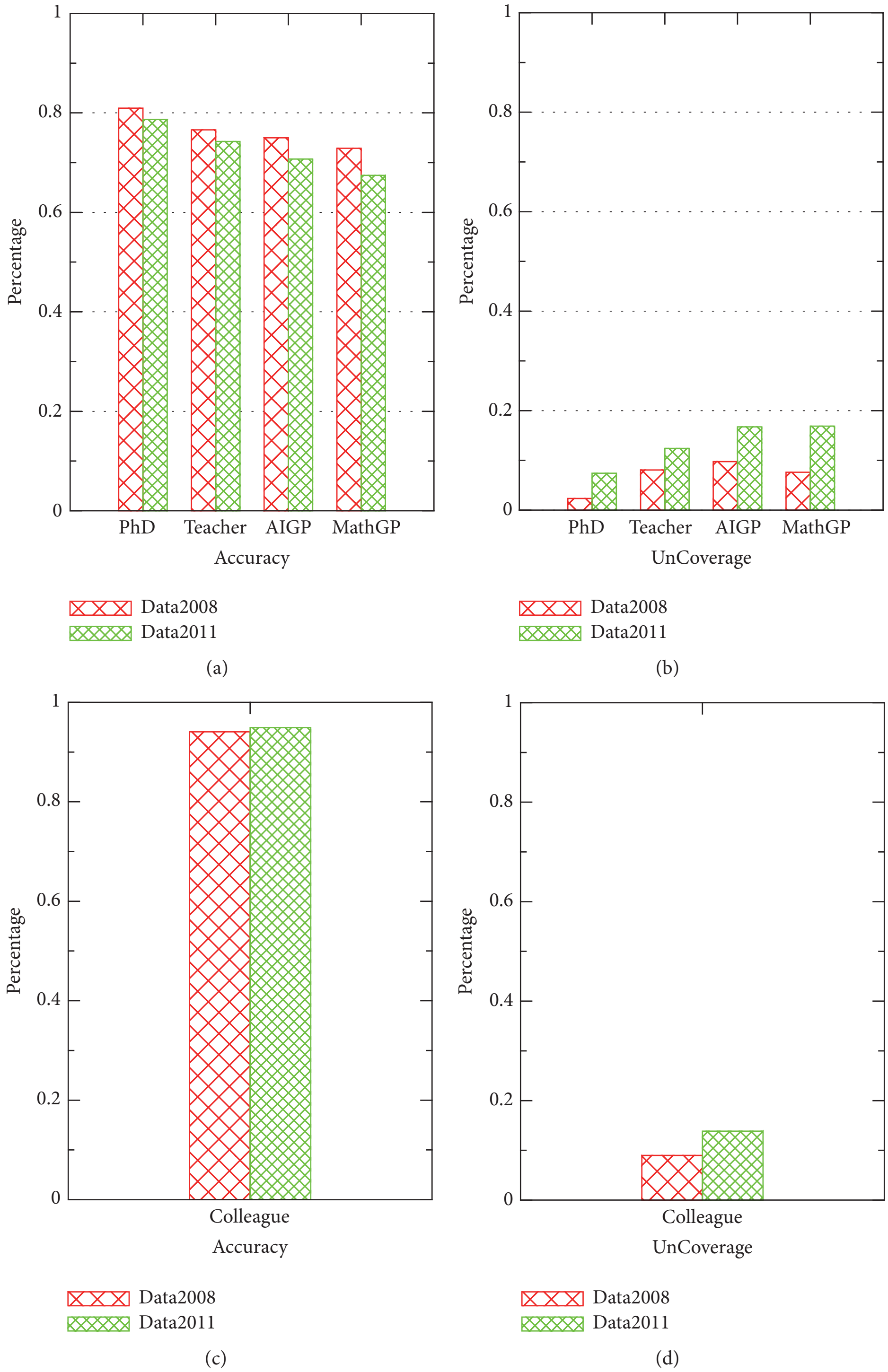
From Figure 4, we find that the largest negative deviation is 6 years and the largest positive deviation is 15 years. Most of the estimated end years lag behind the corresponding labeled end year, that is, the student's graduation year. This is because it is common that a paper is published after the year in which it was written. We also check the advisor-student pairs which have much larger deviation of t_{ij}. Martin Horauer graduated in 2004, but the estimated t_{ij} is 1998. After checking his publication history manually in DBLP, we find that he published two papers with his advisor in 1998 and 1997, respectively. Before he graduated, he did not publish any other paper. On the other side, some students continue to publish their papers with their advisor after they graduated, so the estimated t_{ij} by AMC lags behind the labeled t_{ij}. In these situations, it is impossible to estimate t_{ij} correctly merely based on the publication history.

5. Related Works

Much research has been devoted to Relation Mining [13] and Relational Learning [14]. Those works mainly employ language processing and text mining technique on text data, while our identification method is based on the network topology data other than text data. Another related research branch is link prediction in social networks. Liben-Nowell and Kleinberg [15] explore the unsupervised methods for link prediction in social networks. Backstrom and Leskovec [16] propose a supervised random walk algorithm to estimate the strength of social links. Dong et al. [17] propose a ranking factor graph model for predicting links in social networks, which effectively improves the predictive performance. Leskovec et al. [18] employ a logistic regression model to predict positive and negative links, where the positive links indicate the relationships such as friendship, while negative links indicate opposition. Zhang et al. [19] study the problem of predicting multiple types of links simultaneously for a new LBSN across partially aligned LBSNs, and they [20] also propose a supervised cross aligned networks link prediction with personalized sampling to solve the social link prediction problem for new users. Canfora et al. [21] propose an approach aimed at identifying and recommending mentors in software projects by mining data from mailing lists and versioning systems. Kushwah and Manjhvar [22] summarize recent growth about link prediction algorithms and survey all the prevailing link prediction techniques. However, these works focus on the presence of social relationships, instead of the type classification.

As described in introduction, there are also several works that focus on type classification of relationships hidden in social networks [2, 3]. These two works focus on special domains which are different from our work. Several works or projects have been conducted on the identification or maintenance of types of relationships in research networks. Such projects include the Mathematics Genealogy Project [23], the AI Genealogy Project [24], and the Software Engineering Academic Genealogy [25]. These projects usually ask volunteers to label the advisor or student information for various research fields. However, these methods heavily rely on manual efforts, which significantly limit the amount of data. To overcome the shortcoming of the above manual methods, some research works [5–7] employ the factor graph to infer the type of social relationships automatically. However, these three proposed algorithms are based on factor graph and are computation-intensive.

Some related works about relationship analysis are also studied. Joy et al. [26] examine how students and advisors approach the mutual selection and pairing process, with specific focus on factors influencing the decisions. Tuesta et al. [27] conduct an exploratory study on the advisor-student



(a)

(b)

(c)

(d)

Figure 3: Identification results with different datasets. (a) Accuracy on positive samples. (b) UnCoverage on positive samples. (c) Accuracy on negative samples. (d) UnCoverage on negative samples.

relationship for the researchers who are involved in the area of Exact and Earth Sciences in Brazil. Wan et al. [28] study the problem of finding family groups in the field of civil aviation and propose a family group detection method based on passenger social networks. Lo et al. [29] mine direct antagonistic communities within the signed networks. Noor et al. [30] identify group of people, that is, criminal, terrorist, or friends, by analyzing patterns and performing correlations across different social networks. Lin et al. [31] explore the relationship between value attribution and information source use of 17 Chinese business managers during their knowledge management strategic decision-making. Gayo-Avello et al. [32]

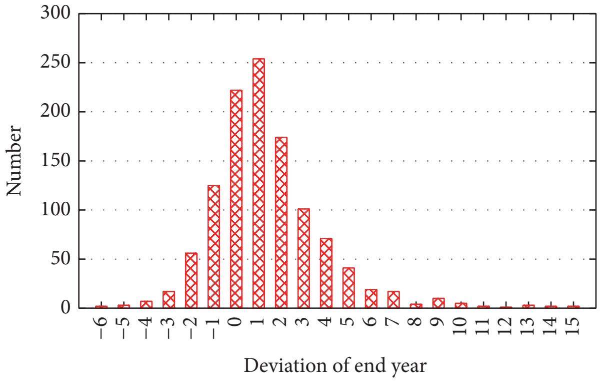


Figure 4: Distribution of deviation of t_{ij}.

offer a comprehensive survey of feasible algorithms for ranking users in social networks. These works are helpful to our work, but their research scopes are different.

6. Conclusion

In this paper, we study the problem of advisor-student relationship identification from publication network. We define this problem in an unsupervised framework and propose an approximate *MaxConfidence*-based algorithm, named AMC, to solve it. We first count the number of papers that two authors coauthored and the number of papers authored by each author, respectively. Next, we employ *AMC* measure to determine the end year and then calculate the relationship score of advising relation. Lastly, we rank every author's potential advisors according to the estimated relationship score and select the one who has the maximum score to be the advisor. The experiment results validate the effectiveness of AMC.

The relationship discovery can give us an opportunity to have a better understanding of the social network. As a research direction in social network analysis, it has attracted many researchers' attention. In the future, the inference performance of AMC still has much space to improve, for example, decrease in the number of name ambiguity or increase in inference accuracy in cases where the information is scarce. In addition, it would be also interesting to investigate how the inference algorithm can be used in other online social networks. How the inferred results benefit applications is also considered in our future work.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Chi Wang from UIUC for sharing his datasets. This research was supported in part by Shaanxi Provincial Natural Science Foundation Research under Grant no. 2014JM2-6104.

References

[1] W. Tang, H. Zhuang, and J. Tang, "Learning to infer social ties in large networks," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 381–397, Athens, Greece, September 2011.

[2] C. P. Diehl, G. Namata, and L. Getoor, "Relationship identification for social network discovery," in *Proceedings of the 22nd National Conference on Artificial Intelligence*, pp. 546–552, Vancouver, Canada, 2007.

[3] N. Eagle, A. S. Pentland, and D. Lazer, "Mobile phone data for inferring social network structure," in *Social Computing, Behavioral Modeling, and Prediction*, pp. 79–88, Springer, New York, NY, USA, 2008.

[4] W. Wang, J. Liu, S. Yu, C. Zhang, Z. Xu, and F. Xia, "Mining advisor-advisee relationships in scholarly big data: a deep learning approach," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pp. 209–210, Newark, NJ, USA, June 2016.

[5] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, and Y. Yu, "Mining advisor-advisee relationships from research publication networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 203–212, ACM, Washington, DC, USA, July 2010.

[6] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogeneous networks," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM '12)*, pp. 743–752, Seattle, Wash, USA, February 2012.

[7] H. Zhuang, J. Tang, W. Tang, T. Lou, A. Chin, and X. Wang, "Actively learning to infer social ties," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 270–297, 2012.

[8] J. Tang, T. Lou, J. Kleinberg, and S. Wu, "Transfer link prediction across heterogeneous social networks," *ACM Transactions on Information Systems*, vol. 9, no. 4, article 43, 2010.

[9] J. Tang, T. Lou, J. Kleinberg, and S. Wu, "Transfer learning to infer social ties across heterogeneous networks," *ACM Transactions on Information Systems*, vol. 34, no. 2, article 7, 2016.

[10] J. He, H. Liu, R. Y. K. Lau, and J. He, "Relationship identification across heterogeneous online social networks," *Computational Intelligence*, 2016.

[11] T. Wu, Y. Chen, and J. Han, "Re-examination of interestingness measures in pattern mining: a unified framework," *Data Mining and Knowledge Discovery*, vol. 21, no. 3, pp. 371–397, 2010.

[12] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, "A brief survey of automatic methods for author name disambiguation," *SIGMOD Record*, vol. 41, no. 2, pp. 15–26, 2012.

[13] B. Coppola, A. Moschitti, and D. Pighin, "Generalized framework for syntax-based relation mining," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)*, pp. 153–162, IEEE, Pisa, Italy, December 2008.

[14] L. Getoor and B. Taskar, Eds., *Introduction to statistical relational learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2007.

[15] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[16] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM '11)*, pp. 635–644, Hong Kong, China, February 2011.

[17] Y. Dong, J. Tang, S. Wu et al., "Link prediction and recommendation across heterogeneous social networks," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM '12)*, pp. 181–190, Brussels, Belgium, December 2012.

[18] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 641–650, Raleigh, NC, USA, April 2010.

[19] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*, pp. 303–312, New York, NY, USA, February 2014.

[20] J. Zhang, X. Kong, and P. S. Yu, "Predicting social links for new users across aligned heterogeneous social networks," in *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM '13)*, pp. 1289–1294, December 2013.

[21] G. Canfora, M. Di Penta, R. Oliveto, and S. Panichella, "Who is going to mentor newcomers in open source projects?" in *Proceedings of the 20th ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE '12)*, Cary, NC, USA, November 2012.

[22] A. K. Kushwah and A. K. Manjhvar, "A review on link prediction in social network," *International Journal of Grid and Distributed Computing*, vol. 9, no. 2, pp. 43–50, 2016.

[23] H. B. Coonce, "Computer science and the mathematics genealogy project," *ACM SIGACT News*, vol. 35, no. 4, p. 117, 2004.

[24] K. Ben, "Artificial intelligence genealogy project," *AI Communications*, vol. 19, no. 1, p. 91, 2006.

[25] Software Engineering Academic Genealogy, http://taoxie.cs.illinois.edu/sefamily.htm.

[26] S. Joy, X. F. Liang, D. Bilimoria, and S. Perry, "Doctoral advisor-advisee pairing in STEM fields: selection criteria and impact of faculty, student and departmental factors," *International Journal of Doctoral Studies*, vol. 10, pp. 343–363, 2015.

[27] E. F. Tuesta, K. V. Delgado, R. Mugnaini, L. A. Digiampietri, J. P. Mena-Chalco, and J. J. Pérez-Alcázar, "Analysis of an advisor-advisee relationship: an exploratory study of the area of Exact and Earth Sciences in Brazil," *PLoS ONE*, vol. 10, no. 5, Article ID e0129065, 2015.

[28] H.-Y. Wan, Z.-W. Wang, Y.-F. Lin, X.-G. Jia, and Y.-W. Zhou, "Discovering Family Groups in Passenger Social Networks," *Journal of Computer Science and Technology*, vol. 30, no. 5, pp. 1141–1153, 2015.

[29] D. Lo, D. Surian, P. K. Prasetyo, K. Zhang, and E.-P. Lim, "Mining direct antagonistic communities in signed social networks," *Information Processing and Management*, vol. 49, no. 4, pp. 773–791, 2013.

[30] F. Noor, A. Shah, and S. A. Khan, "Relation mining using cross correlation of multi domain social networks," in *Proceedings of the SAI Intelligent Systems Conference (IntelliSys '15)*, pp. 898–903, IEEE, London, UK, November 2015.

[31] Y. Lin, C. Cole, and K. Dalkir, "The relationship between perceived value and information source use during KM strategic decision-making: a study of 17 Chinese business managers," *Information Processing and Management*, vol. 50, no. 1, pp. 156–174, 2014.

[32] D. Gayo-Avello, "Nepotistic relationships in Twitter and their impact on rank prestige algorithms," *Information Processing and Management*, vol. 49, no. 6, pp. 1250–1280, 2013.