Searching for Rising Stars in Bibliography Networks

Xiao-Li Li, Chuan Sheng Foo, Kar Leong Tew, See-Kiong Ng
Institute for Infocomm Research, Singapore 138632
{xlli, csfoo, kltew, skng}@i2r.a-star.edu.sg

Abstract. Identifying the rising stars is an important but difficult human resource exercise in all organizations. Rising stars are those who currently have relatively low profiles but may eventually emerge as prominent contributors to the organizations. In this paper, we investigate the problem of identifying rising stars in research communities by mining the social networks of researchers in terms of their co-authorship relationships. We propose a novel PubRank algorithm to mine the evolving links in the social network of researchers modeled by the bibliography network. Our method takes into account the mutual influence of various players in the network, the quality of their publications, as well as the dynamic features of the network over time. Experimental results show that PubRank algorithm can be used to effectively mine the bibliography networks to search for rising stars in the research communities.

Keywords: Rising Stars, Social Network Mining, Bibliography Networks.

1 Introduction

Many organizations are concerned with identifying “rising stars” — those who have relatively low profiles currently but who may subsequently emerge as prominent contributors to their organizations. However, there has been little work on this important task. In this paper, we investigate the possibility of discovering such rising stars from the social networks of researchers constructed using interactions such as research collaborations.

Most of the related social network mining research has focused on discovering groups or communities from social networks [1-4] and on the study of how these communities grow, overlap and change over time [5-7]. In this work, we consider the problem of detecting individual “stars” who rise above their peers over time in the evolving social networks that profile the underlying landscape of mutual influence. In academia and research institutions, it is possible to model the social network of researchers by the bibliography network constructed from their publications. In such a network, the nodes represent individual researchers, while the links denote co-author relationships. Although not explicitly represented in the network, the underlying unit of social interaction is the publication. Each publication signifies a collaborative relationship amongst a certain group of researchers — a link exists between two researchers as long as they have a joint paper together at that point in time. Since each publication has a time stamp, the bibliography network is a dynamic one, constantly changing as the researchers work with different groups of researchers as time goes on.

From such a bibliography network, we aim to discover “rising stars” (nodes in the network) who currently have relatively low profiles but may eventually emerge as prominent researchers. To do so, we need to consider the following factors: 1) *The mutual influence among researchers in the network*. For example, a junior researcher who is able to influence the work of his seniors and effectively collaborate with them, leveraging on their expertise, is far more likely to succeed in a research career. We will model the degree of mutual influence using a novel link weighting strategy. Since each researcher involved in a collaboration may have different contributions to the work, the network is modeled as a bi-directional graph with potentially unequal weights for a link depending on its direction. 2) *The track record of a researcher.* We can measure this in terms of the average quality of the researcher’s current publications. A researcher who publishes in top-tier journals/conferences is more likely to be an influential researcher than another who publishes at less significant venues. This can be accounted for by placing different weights on different nodes in the network model. 3) *The chronological changes of the networks*. Each researcher may work with different groups of people at different points in time. A researcher who can build up a strong collaborative network more rapidly than others is more likely to become a rising star. This means that our network model needs to be time-stamped based on the publication data.

In this work, we design a novel PubRank algorithm to mine rising stars from bibliography networks by incorporating the factors described above. We will show that our method works well on DBLP Computer Science Bibliography data. The main contributions of our work are summarized as follows:

- A novel social network mining algorithm PubRank has been designed to analyze the dynamic collaborative landscape of researchers based on bibliography networks.
- Our algorithm derives information from the out-links of nodes, which is fundamentally different from many other related node analysis algorithms, which use information from the in-links.
- The algorithm has been applied on more than a million computer science publications from all over the world to detect rising stars with promising results.
- Our technique is potentially useful for academia and research institutions in their recruitment and grooming of junior researchers in their organizations. It may also be useful to fresh PhDs and postdocs for selecting promising supervisors. Finally, it can be useful for tracking one’s relative performance in the research community, and for deciding whom to collaborate (more) with.

The rest of the paper is organized as follows. First, we provide an overview of the related work in social network mining in Section 2. Then, we present our proposed PubRank algorithm for mining rising stars from publication data in Section 3. This is followed by the presentation of the experimental results from applying our proposed technique to the DBLP bibliography data in Section 4. We conclude the paper with some discussions on possible future research directions in Section 5.

2 Related Work

While link- or network-based analysis has been studied in social network analysis (SNA) and web information retrieval, there has not been much cross-fertilization of ideas between the different research communities until the recent years. The subject has since become an exciting and rapidly expanding area [9]. This has resulted in a number of network mining algorithms and applications. In general, the methods can be categorized by the various levels at which they mine the networks, namely, at the node level, link level, and community or network level.

For node level network mining, the most well known techniques are the PageRank [10] and HITS algorithms [11] proposed for the Web information retrieval domain. The PageRank algorithm models Web surfing as a random walk. A web surfer randomly selects/follows links and occasionally jumps to a new web page to start another traversal of the link structure. As our PubRank algorithm is related to the PageRank method, we will describe the details of the PageRank measure in Section 3 where we will also present our algorithm and highlight the differences. Our algorithm incorporates a twist on the original PageRank algorithm in using information from the out-links of a node as opposed to the in-links, as is done in most node analysis algorithms. The HITS algorithm models the Web as one comprising two types of key web pages — hubs and authorities. Hubs are web pages that link to many authoritative pages, whereas authorities are web pages that are linked to by many hubs. Given a web page, its hub and authority scores are computed by an iterative algorithm that updates these scores based on the scores of pages in its immediate neighborhood. Other more recent methods for ranking entities in networks based on the relations between the entities are presented in [12] and [13].

Entity resolution is another important research problem at the node level. The problem here is to detect which references in the data refer to the same entity. In our context, it will be to resolve different researchers who have the same name. Current approaches for the entity resolution problem have been to combine the network structures with a feature–based method to improve accuracy [14] [15]. However, entity resolution has remained a challenging problem. For the sake of brevity, we will not consider it within the scope of this paper.

At the link level, a typical network mining problem is link prediction. This involves predicting missing links between entities based on the attributes of the entities and/or other observed links. The network mining techniques proposed so far are based on graph proximity measures, attribute information, and structured logistic regression models [16] [17].

For network mining at the community level, a widely researched problem is community detection. For example, Palla *et al.* showed that meaningful overlapping community structures can be detected from different real world networks, such as collaboration networks, word-association networks and protein interaction networks [3]. Community detection in network mining involves clustering the nodes in the graph into (possibly overlapping) groups that share some common characteristics. The block modeling method was a classical SNA method for this problem [2]. More recently, spectral graph partitioning methods have also been employed to detect the groups by identifying an approximately minimal set of links from the input graph to achieve a given number of groups [1]. A related problem is subnetwork or network motif discovery, which involves finding interesting and commonly occurring subnetworks in a set of networks. Several approaches have been proposed to tackle this computationally intensive problem by exploiting the a-priori property from frequent item set mining [18] [4] [19].

As mentioned earlier, real world social networks such as the researchers' bibliography networks are dynamic networks that evolve over time. To take this into consideration, network mining methods must analyze networks at the network level [5] [6] [7]. Many interesting problems have been recently explored. For example, the work in [5] investigates communities that grow rapidly and how the overlaps between pairs of communities change over time, while the work in [6] tries to discover what are the "normal" growth patterns in social, technological and information networks. [7] proposes a tractable model for information diffusion in social networks.

In this paper, we aim to detect the rising stars by mining the bibliography networks. To our best knowledge, this is the first attempt to discover potential star researchers or hidden talents using information in a bibliography network. The work by Mohan [20] for detecting "nurturers" from association networks seems related to our rising star problem. However, mining for nurturers is a different problem from mining for rising stars — the nurturers correspond to researchers who are already influential, whereas mining for rising stars involves detecting researchers who are yet to be have made their mark. Moreover, the method is targeted for Web information retrieval. Another notable SNA application is the work by Chau *et al.*, which involves detecting fraudulent personalities in networks of online auctioneers [21]. Again, this is a different problem from rising star detection.

3 The Proposed Technique

We present our proposed technique in this section. First, we describe the construction of a directed, weighted bibliography network from bibliography data to model the social relationships among researchers in Section 3.1. Next, we describe the computation of node weights, which incorporate information regarding the quality of a researcher's publications, in Section 3.2. We then define a novel PubRank score that models the propagation of mutual influence among researchers in the constructed network and describe its relationship to the PageRank algorithm [10] in Section 3.3. We briefly discuss the convergence of the power method used to compute the PubRank score in Section 3.4. Section 3.5 presents our approach to account for the evolution of the network over time in order to detect rising stars. Finally, the detailed PubRank algorithm is presented in Section 3.6.

3.1 Constructing the bibliography network

We define a bibliography network to be a directed, weighted network $G = (V, E)$, where the node set V consists of authors in the network and the edge set E describes all co-author relationships. More formally, $V = \{ v_i \mid v_i$ is a author in publication data$\}$ and $E = \{(v_i , v_j) \mid$ *co-pub* $(v_i , v_j) > 0,\ v_i,\ v_j \in V\}$, where *co-pub*$(v_i , v_j)$ denotes the number of publications in which authors v_i and v_j are co-authors. It is important to note that in this directed network *G*, edge (v_i , v_j) and edge (v_j , v_i) are two different edges with potentially unequal weights; the weight w(v_i , v_j) represents the influence of author v_i on v_j while the weight w(v_j , v_i) represents the influence of author v_j on v_i.

When two authors v_i and v_j co-author a publication, there is mutual influence between them as the collaboration is typically beneficial to both parties. Our proposed PubRank algorithm models the mutual influence between the two co-authors by weighting the edges in the network G as follows:

$$w(v_i, v_j) = \frac{co\_pub(v_i, v_j)}{\sum_{k=1}^{|V|} co\_pub(v_j, v_k)} \qquad (1)$$

where *co_pub*(v_i, v_j) is the number of publications that v_i and v_j have co-authored. The denominator is also the total number of publications that v_j has (co-)authored.

The proposed weighting scheme uses the number of publications co-authored by each pair of researchers as a proxy for the strength of their collaboration relationship. Researchers are then modeled to influence each other according to the strength of this relationship. An expert v_i will tend to influence a junior researcher v_j more than the junior influences the expert. This is modeled in our scheme with the weight w(v_i , v_j) typically being bigger than w(v_j , v_i) as the junior researcher v_j is likely to have fewer total publications, resulting in a smaller denominator for w(v_i, v_j).

We provide an example to illustrate the use of equation (1) in Figure 1. In the figure, node (researcher) 1 has four direct neighbors (co-authors) 2, 3, 4 and 5 while node 2 has neighbors 1 and 6. We also marked the number of co-authored publications between any two nodes. Here, node 2 has 5 publications where 4 are with node 1 and 1 with node 6. Thus, w(1,2)=4/(4+1)=0.80 and w(6,2)=1/(4+1)=0.20, indicating that node 1 has bigger influence to node 2 than node 6. Similarly, for node 1, we can compute the weights w(2,1)=4/18=0.22, w(3,1)=5/18=0.28, w(4,1)=8/18=0.44 and w(5,1)=1/18=0.06. Note that w(2,1)+w(3,1)+w(4,1)+w(5,1)=1, and w(1,2) > w(2,1) since node 1 is likely to be a more established researcher and therefore has a higher influence on node 2, who is likely to be a junior researcher and publishes papers mainly with node 1. Our model is therefore able to take into account the observation that a junior researcher or student often publishes most of his or her papers with a supervisor or professor, whereas an established researcher is likely to co-author papers with a more diverse set of researchers, including colleagues, collaborators and students.

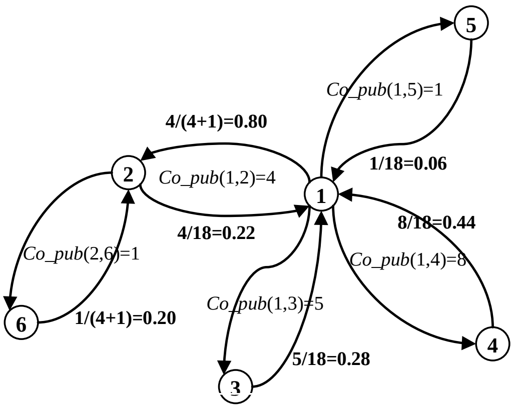

Fig. 1. Assigning weights to the edges in a bi-directed bibliography network.

3.2 Accounting for the quality of publications: assigning node weights

In the previous step, we have used the number of co-authored publications between pairs of researchers to assign weights to the directed edges in the bibliography network. This allows us to model the mutual influence between the authors (nodes). However, the reputation and impact of a researcher is also decided by the quality of his/her work. Specifically, if an author has published most of his/her papers in top-tier conferences and journals, then he/she is more likely to be a well-recognized domain expert. As such, his/her work will tend to receive more attention from the community and thus exert more influence on the field. There is thus a need to account for the quality of a researcher's work in our algorithm.

We have taken the approach of incorporating this information by assigning node weights using the quality of a researcher's publications. One common approach to evaluating the quality of a paper is by its citation count. However, this measure is biased towards earlier publications because more recently published articles need time to accumulate citations [20]. As we are trying to discover rising stars, these junior researchers are unlikely to have many highly cited papers. Therefore, instead of using citation count to gauge the quality of a paper, we opted for an alternative measure that is based on the prestige of its publication venue. Numerous ranking schemas (e.g., http://citeseer.ist.psu.edu/impact.html, http://www.cs-conference-ranking.org/, http://scientific.thomsonreuters.com/products/jcr/) are available for this purpose. They rank conferences and journals according to their acceptance ratio, publication of exceptional results, participation by famous researchers in the conference and the members of the program committee [22]. Using such information, conferences and journals can be approximately ranked by their quality. A commonly used system is as follows: rank 1 (premium), rank 2 (leading), rank 3 (reputable) and unranked [22].

Given a paper, we compute a measure of its quality based on the rank of the corresponding conference or journal where it was published. Then, given an author v_i who has a publication set P, we define his/her publication quality score $\lambda(v_i)$ as follows:

$$\lambda(v_i) = \frac{1}{|P|} * \sum_{i=1}^{|P|} \frac{1}{\alpha^{r(pub_i)-1}}, \tag{2}$$

where pub_i is the i-th publication, $r(pub_i)$ is the rank of publication pub_i, and α (0 < α < 1) is a damping factor designed so that lower ranked publications have lower scores. The larger $\lambda(v_i)$ is, the higher the average quality of papers published by researcher v_i.

3.3 Propagating influence in the bibliography network

We have shown how to construct a directed network that is both edge- and node-weighted to model the social interactions between researchers via co-authorship. The benefit of having a co-author is mutual — a young researcher will stand to gain by working with a more experienced and established collaborator, while the experienced researcher is far more productive by teaming up with like-minded researchers (both experts and promising novices) to do good work. Naturally, new researchers would tend to have fewer collaborators compared to their experienced counterparts who are bound to be more networked and selective in whom they collaborate with. As the mutual benefit encourages researchers to collaborate, the underlying landscape of mutual influence changes continually as researchers interact (i.e. collaborate) and grow in stature. This feedback nature has also been famously observed in another real-world network, namely the “social network” of co-referencing web-pages. The PageRank algorithm [10] is the best-known web-page ranking algorithm used by the successful Google search engine. It ranks web-pages using a scoring scheme based on the hyperlinks among the web pages. The PageRank score of a webpage is defined as follows:

$$PR(p_i) = \frac{1-d}{N} + d * \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{3}$$

where p_i ($i = 1, 2, \ldots, |N|$) is the webpage under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, N is the total number of pages, and the parameter d is usually set to 0.85 for various practical considerations.

The PageRank of a page p_i is defined in terms of the PageRanks of all the pages that link to a page p_i. Specifically, the summation term in formula (3) reflects the sum of the PageRanks of all pages (e.g., p_j) that link to a page p_i divided by the number of outgoing links from page p_j, the intuition being that the more outgoing links a page has, the less important each link is. Thus the 'value' of each link is uniformly distributed amongst all links originating from a particular page. The (1-d)/N term at the beginning ensures that the adjacency matrices used will be stochastic matrices, so the resulting PageRanks satisfy a property that the sum of all web pages' PageRanks will be 1. It also means that if a page has no links to it, it will still get a small PageRank of 0.15/|N| [10].

The PageRank of a set of web pages can be calculated using the *power method*, a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. Theoretically, the power method will converge to the true eigenvector given sufficient iterations [10][23].

In this work, we adapted the PageRank algorithm to compute a similar score for each node based on the propagation of influence in the bibliography network. Note that the score of a page p_i in the PageRank algorithm is based solely on the number of outgoing links of p_j that link to a page p_i. In our scenario, we consider both mutual influence between authors (equation 1) and the quality of each author’s publications (equation 2) to compute a similar PubRank score for each author (node):

$$PubRank(p_i) = \frac{1-d}{N} + d * \sum_{j=1}^{|V|} \frac{w(p_i, p_j) * \lambda(p_i) * PubRank(p_j)}{\sum_{k=1}^{|V|} w(p_k, p_j) * \lambda(p_k)} \tag{4}$$

Using the PubRank score from equation (4), we can model the propagation of mutual influence amongst researchers. A researcher with many excellent papers will have a larger influence on fellow co-authors. Note that the influence may not be from direct neighbors as all the authors in the network also indirectly affect others.

Closer examination of equation (4) will reveal that unlike the PageRank score, in which the PageRank of a node is influenced by the scores of *nodes that link to it*, in the computation of our PubRank score, the PubRank of a node is dependent on the *nodes to which it links to*. In other words, unlike in the PageRank algorithm and most other link analysis algorithms which use *in-links* to derive information about a node, our algorithm uses a node's *out-links* to compute its score. This innovation on the regular approach in node analysis algorithms reflects the reality of the situation – if a researcher has high quality publications, and is able to contribute to the work of other influential researchers, then he/she is likely to be a rising star.

3.4 Convergence of the PubRank computation

In this section, we discuss the convergence of the power method when used to compute PubRank scores. We first express the computation of the PubRank score as an eigenvalue problem and show that the matrix involved is stochastic. This property ensures that the power method converges to the desired PubRank score. In practice, the algorithm is fast due to the sparseness of the bibliography network. In our experiments, it converges in approximately 50 iterations, which takes a few minutes on the entire dataset of over a million publications.

Define the $|V|$ by $|V|$ weight matrix W such that $W_{ij} = w(p_i, p_j)\lambda(p_i) / Z_j$, where $Z_j = \sum_{k=1}^{|V|} w(p_k, p_j)\lambda(p_k)$ is a normalization constant. By construction, we see that the entries in each column of W sum to 1, so W is a column stochastic matrix. We may then rewrite the PubRank computation as

$$u = (1\text{-}d)E + dWu$$

where u is the vector of PubRanks and E is the vector which has $1/|V|$ for all entries. Thus, in the power method computation

$$u^{(t+1)} = (1\text{-}d)E + dWu^{(t)}$$

the sequence of iterates $u^{(t)}$ converges in the limit to the PubRank score u^*, which is also given explicitly by $u^* = (1\text{-}d)(I - dW)^{-1}E$, where I is the $|V|$ by $|V|$ identity matrix.

3.5 Discovering rising stars from the evolving networks

Up till this point, we have not accounted for a very important characteristic of a bibliography network, which is its evolution through time. As mentioned, a bibliography network is clearly not a static network and it changes dynamically over time. As young researchers mature, they will also slowly make a foothold in their respective domains and begin to enrich their own social networks with their own influence, nurturing new young researchers.

In this work, we consider each calendar year as the unit of time steps. Given an author v_i ($i = 1, 2, \ldots, n$) and his/her historical PubRank scores $p(v_i, t_1)$, $p(v_i, t_2)$, …, $p(v_i, t_m)$ at the time point (calendar year) $t_1, t_2, t_3, \ldots, t_m$, we predict if he/she will be a rising star in k years' time t_{m+k} based on his/her performance in the past m years. For this work, we adopt the hypothesis that if a researcher v_i demonstrates an increase in his/her annual PubRank scores that is significantly larger than those of an average researcher, he/she will probably do very well in the coming years. In addition, we also require $p(v_i, t_1)$, has to be lower than the average PubRank score of all researchers at

t_1, i.e. $p(v_i, t_1) < \frac{1}{|V|}\sum_{k=1}^{|V|} p(v_k, t_1)$. This allows us to search for the "hidden" rising stars.

We use a linear regression model to identify the rising stars, which is formulated as follows:

$$\hat{p} = k \times t + b \tag{5}$$

To minimize the residual sum of squares $\sum_{j=1}^{m}(p(v_i, t_j) - \hat{p}(v_i, t_j))^2$ (i.e. best fit the historical data), the gradient k can be computed by using the equation (6):

$$k_i = (\sum_{j=1}^{m} t_j\, p(v_i, t_j) - \frac{\sum_{j=1}^{m} t_j \sum_{j=1}^{m} p(v_i, t_j)}{m}) / \sum_{j=1}^{m}(t_j^{\,2} - \frac{(\sum_{j=1}^{m} t_j)^2}{m}) \tag{6}$$

For each researcher v_i, we compute the gradient k_i based on the PubRank scores over time points t_1 *to* t_m. A researcher v_i with a large positive gradient k_i means his/her PubRank scores have increased significantly and thus likely to shine in near future.

Once we have computed the gradient distribution k_i (i = 1, 2, …, n), we assess the significance of the gradient by computing its Z-score:

$$z(v_i) = \frac{k_i - \mu}{\sigma} \tag{7}$$

Where μ and σ are the mean and standard deviation respectively for the gradient distribution. Assuming that the gradients of the researchers have a Gaussian distribution, a critical region typically covers 10% of the area (the probability distribution) in the tail of the distribution curve. Thus, if $z(v_i) \geq 1.282$, we regard it as statistically significant and v_i will be predicted as a rising star.

3.6 The overall PubRank algorithm

Our overall algorithm for detecting the rising stars is shown in Figure 2.

```
Input: publication set $P$, time start point $t_1$ and end point $t_m$
Output: rising star set $RS$
$RS = \Phi$;
Temporal point set $T = \{t_1, t_2, t_3, \ldots, t_m\}$;
For each time point $t_i \in T$
    $P_i = \{p_i \mid time(p_i) <= t_i, p_i \in P\}$;
    $A_i = \{v_i / pub(v_i, P_i) > 0\}$;
    For all the publication $p \in P_i$
        If $v_i$ ($v_i \in A_i$) and $v_j$ ($v_j \in A_i$) are co-authors at time point $t_i$ in set $P_i$;
            $E_i = E_i \cup \{(v_i, v_j)\}$ ;
            $E_i = E_i \cup \{(v_j, v_i)\}$;
            Assign weights $w(v_i, v_j)$ and $w(v_j, v_i)$ using equation (1);
    For each $v_i \in A_i$
        Compute the publication quality $\lambda(v_i)$ for $v_i$ using formula (2);
    Run our iterative algorithm (equation 4) for all the author $v_i$ ($v_i \in A_i$) output a
    influence score PubRank($v_i$) ;
```

```
14. For each $v_i \in \bigcup_{i=1}^{m} A_i$
15.     Compute the gradient $g(v_i)$ using formula (6) ;
16. Compute the mean $\mu$ and standard deviation $\sigma$ for the gradient distribution;
17. For each $v_i \in \bigcup_{i=1}^{m} A_i$
18.     If $(p(v_i, t_1) < \frac{1}{|\bigcup_{i=1}^{m} A_i|} \sum_{k=1}^{|\bigcup_{i=1}^{m} A_i|} p(v_k, t_1))$
19.         Compute its $Z$-score using equation (7);
20.         If $z(v_i) \geq 1.282$
21.             $RS = RS \bigcup \{v_i\}$.
```

Fig 2. Overall PubRank algorithm to mine the rising stars from the bibliography network.

After initializing the rising star set RS and the *temporal point set* T in steps 1 and 2, steps 3 to 13 assign a PubRank score for each author. Steps 6 to 10 construct the directional and weighted network; steps 11 to 12 compute publication quality scores; step 13 runs our adapted iterative propagation step. Steps 14 to 15 compute the gradient for each author. After we have computed the mean and standard deviation in step 16, step 18 makes sure that the rising stars' score p(v_i, t_1) is lower than the average PubRank score. Finally, we transform the PubRank score into Z-score and identify those authors with significant high Z-score gradients as rising stars.

4 Experimental Evaluation

We tested our proposed technique by mining for rising stars from large bibliography networks. In our experiments, we used publication data from the Digital Bibliography and Library Project (DBLP). The DBLP database provides bibliography information on major computer science journals and conferences (http://www.informatik.uni-trier.de/~ley/db/). DBLP currently lists more than one million articles; each article record contains the author names, title, conference or journal name, year of publication, as well as other bibliographic information. For our work, we used only the author names, conference or journal names, and year of publication.

In our first experiment, we use all the DBLP data which spans various Computer Science domains, for example, "Artificial Intelligence", "Information Retrieval", "Databases", "Multimedia" and "Bioinformatics". This large data set with over one million publications tests the scalability of our algorithm. Next, in our second experiment, we evaluate our algorithm on a subset of DBLP data from the Database domain. This is because one is often more interested in the performance of one's peers in the same technical domain than the entire field of computer science. The Database domain was chosen here due to its long pedigree and relevance to our work in data mining. In our experiments the damping factor α is set to 2 which means the publications of rank 1 (premium), rank 2 (leading), rank 3 (reputable) and unranked (see the rank schema [22]) are assigned the weights 1, 1/2, 1/4, 1/8 respectively.

Results on the entire DBLP dataset. First, we used the historical data from 1990 to 1995 to predict the rising stars. Then, we “fast forward” a decade ahead and look at the eventual PubRank scores of our predicted stars in year 2006 to verify if our rising stars have indeed realized their predicted potential.

We normalized the PubRank scores of all researchers using the Z-score measure as described in our method. Out of the 64,752 researchers with high PubRank scores (Z-score > 0), we have identified 4,459 rising stars using our proposed method.

We compare the rising stars with researchers in general. We noticed that on average, the rising stars continued to have significantly higher gradients in the period after 1995 — the average gradient for the rising stars is 0.497 while the average gradient for all researchers is 0 (Z-score property). Our predicted stars have indeed increased their PubRank scores significantly faster than researchers in general (Figure 3). In fact, although the rising stars all started out as relatively unknown researchers in 1990 (with PubRank scores lower than average), their final average Z-score in 2006 were 2.92, which means that they were eventually ranked in the top 1% of all researchers ten years later.

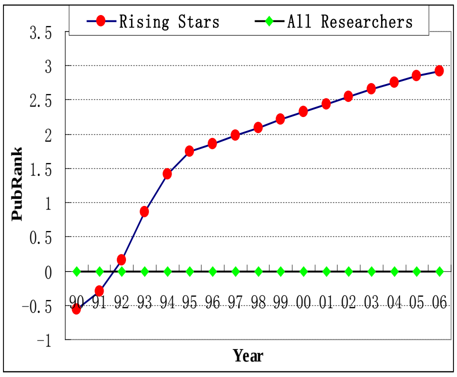

Fig 3. The comparison of the Z-score of rising stars and all researchers across the years.

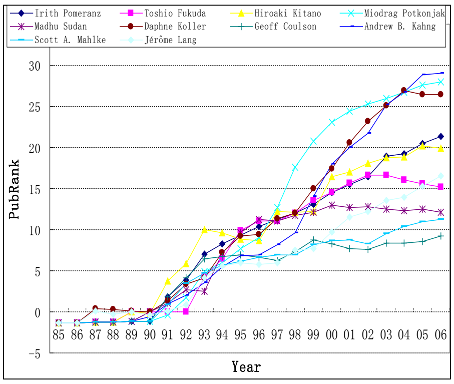

Fig. 4. The increasing trend in PubRank scores for the top ten rising stars from year 1990 to 1995

We further investigated the top ten rising stars and their detailed PubRank scores from 1985 to 2006 (Figure 4). We noticed that these rising stars showed a pattern of continuous increase. On average, the top ten rising stars in 2006 had a PubRank score of 18.90, which is 6.47 times than the average scores for rising stars in general.

Table 1 shows our predicted top 10 rising stars with their gradients (which were computed using their PubRank scores from 1990 to 1995), their starting ranks in 1990, and their final ranks in 2006. On average, the gradient of all top ten rising stars was 1.79. Their average starting ranking (based on their PubRank scores in 1990) was 47922.8, out of the 64,752 authors (the authors will be ranked at 64753 if their earliest publication is later than 1990). In 2006, their average ranking became 265 out of 502,481 authors. This shows that our algorithm has indeed successfully predicted researchers who have risen significantly in the past decade using the bibliography networks.

Another interesting observation in Table 1 is that many rising stars have at least one nurturer and/or strong collaborator. Out of the top ten rising stars, we found that 9 of them co-published papers with researchers with much higher PubRank scores than

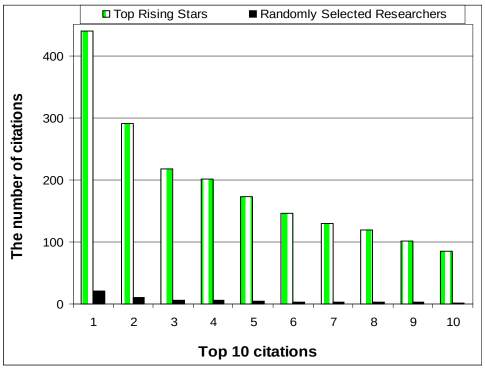

Fig. 5. The comparison of the number of citations between the rising stars and randomly selected researchers

them, suggesting that during the initial stages of their careers, these rising stars were under the supervision of experienced collaborators who probably provided a solid foundation for the stars' future career growth.

Figure 5 further shows the number of citations of the top ten rising stars compared with randomly selected 100 non-rising star researchers using Google Scholar (http://scholar.google.com). We noticed that the rising stars obtained significantly higher citations than the random researchers. In particular, the average citation of the best cited paper of rising stars is 440 compared to only 18.9 for randomly selected researchers. In average, each of 10 rising stars managed to obtain an impressive 1904.2 total citations of their top ten cited papers, which is 340 times than the randomly selected researchers which can manage to get only 5.6 citations of their top 10 cited papers. The significantly higher citations indicate that the rising stars have indeed become well established experts and are highly recognized in their respective domain.

Table 1. Top 10 predicted rising stars and their respective rank in year 1990 and 2006.

Name	**Gradient**	**Author Ranking**		**Possible nurturers and their rank (in 1990)**
		1990	**2006**	
Irith Pomeranz	2.13	64753	85	Sudhakar M. Reddy(67)
Toshio Fukuda	2.09	17992	244	
Hiroaki Kitano	1.91	38330	106	Hideto Tomabechi(5620)
Miodrag Potkonjak	1.89	64753	24	Alice C. Parker(32)
Madhu Sudan	1.89	34896	466	Richard J. Lipton(5) Daniel H. Greene(683)
Daphne Koller	1.85	34901	32	Danny Dolev(7), Amotz Bar-Noy(323), Rudiger Reischuk(681)
Geoff Coulson	1.65	64753	915	Gordon S. Blair(964)
Andrew B. Kahng	1.49	50451	19	C. K. Wong(543), Majid Sarrafzadeh(1613)
Scott A. Mahlke	1.49	64753	568	Wen-mei W. Hwu(696)
Jérôme Lang	1.47	43646	191	Henri Prade(665) Didier Dubois(1156)
Average	**1.79**	**47922.8**	**265**	

As a final test, we ran our PubRank algorithm to mine the rising stars using the publication data from 1950 (1950–1955) to 2002 (2002–2007). In order to validate our predictions, we chose the *h*-index list (http://www.cs.ucla.edu/~palsberg/h-number.html) which is used to quantify the cumulative impact and relevance of an individual's scientific research output. The *h*-index, defined as the number of papers with citation number higher or equal to *h*, is a useful index to characterize the scientific output of a researcher [25]. Out of the 131 researchers with 40 or higher *h*-index score according to Google Scholar, 116 researchers (88.5%) are identified as rising stars by our algorithm across different years.

Results on the Database domain. Here we investigate the effectiveness of our algorithm when considering only publications from the database domain. A list of database conferences is obtained from schema [22]. We retrieved 19474 papers published at these venues from the DBLP data. Our PubRank algorithm is subsequently used to identify the rising stars from 1990 to 1994 (rising stars in year *n* are predicted using historical data from *n-5* to *n-1*). Note that a researcher can be predicted as a rising star in multiple years if their scores are always increasing significantly. To validate the results of our algorithm, we choose the top 20 rising stars for each year from 1990 to 1994. Out of the 100 rising stars, there are 63 unique individuals. Manual evaluation of the achievements of the 63 individuals yields the following results: 1) 43 (or 68.3%) have been appointed full professors at renowned universities. 2) 7 (11.1%) of them are key appointment holders (Founder/President/Directors) at established research laboratories and companies. 3) the remaining 13 are either Associate Professors or hold important positions in industry.

Table 2. Top predicted rising stars from database domains from year 1990 and 1994.

Name	Position	Organization	Awards	Top Citation
Bharat K. Bhargava	Professor	Purdue University	IEEE Technical Achievement Award, IETE Fellow	143
H. V. Jagadish	Professor	University of Michigan, Ann Arbor	ACM Fellow	457
Hamid Pirahesh	Manager	IBM Almaden Research Center	IBM Fellow, IBM Master Inventor	1428
Ming-Syan Chen	Professor	National Taiwan University	ACM Fellow, IEEE Fellow	1260
Philip S. Yu	Professor	UIC	ACM Fellow, IEEE Fellow	1260
Rajeev Rastogi	Director	Bell Labs Research Center, Bangalore	Bell Labs Fellow	1178
Rakesh Agrawal	Head	Microsoft Search Labs	ACM Fellow, IEEE Fellow, a Member of the National Academy of Engineering	6285
Richard R. Muntz	Professor	UCLA	ACM Fellow, IEEE Fellow	1191
Shi-Kuo Chang	Professor	University of Pittsburgh	IEEE fellow	171
Jiawei Han	Professor	UIUC	ACM fellow	6158

Table 2 shows the achievements of a selection of 10 outstanding individuals from the 63 we earlier identified. Their most highly cited publications all have over 100 citations (as found using Google Scholar) and 7 of them have been recognized as ACM and IEEE fellows (or both). The other individuals that we identified (names not listed) also have remarkable achievements such as being appointed editor-in-chief for prestigious journals or winning (10 year) best papers at major database conferences (*e.g.*, SIGMOD, PODS, VLDB, ICDE, KDD, ICDM). Such achievements are clear indicators that they have indeed become the shining stars in the database domain, as we have predicted with our algorithm with publication data more than a decade ago.

5 Conclusions

Rising stars are persons who initially have relatively low profiles but who may eventually become prominent contributors to the organizations. Identifying and recruiting the rising stars of tomorrow is vital for the growth of all organizations. In this paper, we have proposed a novel social network analysis technique to mine the rising stars from the social networks of the research communities based on their bibliography networks. We have developed a novel link weighting strategy to model the mutual influence among the researchers. We designed a node-weighting scheme in our network model to take the track records of each researcher into account. We have devised the PubRank algorithm to propagate the mutual influence in the bibliography network so that we can identify the rising stars by tracking how the underlying mutual influence landscape changes over time.

Comprehensive experimental results showed that our technique can indeed be used to effectively search for rising stars in the research communities. In addition, the fact that we are able to obtain promising results using an algorithm which considers out-links instead of in-links is an interesting twist on existing node analysis algorithms and suggests that such an approach to network mining problems should be further explored.

There are still numerous aspects of our work which can be improved upon. For example, we can take into consideration other information such as the ordering of authors in the publications. In the current study, we have treated all publications with the same author name as being from the same author — this can be problematic for those authors with common names. For future work, we will develop and employ more sophisticated named entity recognition techniques in the construction of the social networks.

References

[1] M. E. J. Newman, "Detecting community structure in networks," *European Physical Journal B*, vol. 38, pp. 321-330, 2004.

[2] S. Wasserman and K. Faust, *Social network analysis: methods and applications*: Cambridge University Press, 1994.

[3] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814-818, 2005.

[4] X. Yan and J. Han, "gSpan: Graph-based substructure pattern mining," in *ACM Conference on Data Mining (KDD)*, 2002.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group Formation in Large Social Networks: Membership, Growth, and Evolution," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, USA, 2006.

[6] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations " in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.

[7] M. Kimura and K. Saito, "Tractable Models for Information Diffusion in Social Networks " in *ECML/PKDD*, 2006.

[8] L. Getoor and C. P. Diehl, "Link Mining: A Survey," *SIGKDD Explorations,* vol. 7, pp. 3-12, 2005.

[9] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," in *Proceedings of the seventh international conference on World Wide Web*, 1999, pp. 107-117.

[10] J. Kleinberg, "Authoritative sources in a hyperlinked enviroment," *Journal of the ACM,* vol. 46, pp. 604-632, 1999.

[11] Z. Zhuang, S. Cucerzan, and C. L. Giles, "Network Flow for Collaborative Ranking," in *ECML/PKDD*, 2006.

[12] A. Agarwal, S. Chakrabarti, and S. Aggarwal, "Learning to Rank Networked Entities," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, USA, 2006.

[13] R. Ananthakrishna, S. Chaudhuri, and V. Ganti, "Eliminating fuzzy duplicates in data warehouses," in *International Conference on Very Large Databases (VLDB)*, Hong Hong, China, 2002.

[14] X. Dong, A. Halevy, and J. M. adhavan, "Reference reconciliation in complex information spaces," in *ACM SIGMOD International Conference on Management of Data*, 2005, pp. 85-96.

[15] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *International Conference on Information and Knowledge Management (CIKM)*, 2003, pp. 556-559.

[16] A. Popescul and L. H. Ungar, "Statistical relational learning for link prediction," in *IJCAI Workshop on learning statistical models from relational data*, 2003.

[17] A. Inokuchi, T. Washio, and H. Motoda, "An Apriori-based algorithm for mining frequent substructures from graph data," in *European Conference on Principles and Practice of Knowledge Discovery and Data Mining*, 2000, pp. 13-23.

[18] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng, "NeMoFinder: Dissecting genome wide protein-protein interactions with repeated and unique network motifs," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, USA, 2006, pp. 106-115.

[19] L. Licamele and L. Getoor, "Social Capital in Friendship-Event Networks," in *ICDM* Hong Kong, 2006.

[20] B. K. Mohan, "Searching association networks for nurturers," *Computer, IEEE Computer Society,* vol. 38, pp. 54-60 2005.

[21] D. H. Chau, S. Pandit, and C. Faloutsos, "Detecting Fraudulent Personalities in Networks of Online Auctioneers " in *ECML/PKDD2006*, 2006.

[22] P. M. Long, T. K. Lee, and J. Jaffar, "Benchmarking Research Performance in Department of Computer Science, School of Computing, NUS " *http://www.comp.nus.edu.sg/~tankl/bench.html*, 1999.

[23] B. Liu, *Web Data Mining* Springer, 2006.

[24] R. Collobert and S. Bengio, "SVMTorch: support vector machines for large-scale regression problems," *The Journal of Machine Learning Research* vol. 1, pp. 143-160, 2001.

[25] J. E. Hirsch, "An index to quantify an individual's scientific research output," *PNAS,* vol. 102, pp. 16569-16572, 2005.