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 propose a novel PubRank algorithm to identify rising stars in research communities by mining the social networks of researchers in terms of their co-authorship relationships. 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 [1-2] and on the study of how these communities grow, overlap and change over time [3]. 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 universities and research institutions, it is possible to model the social network of researchers by the bibliography network constructed from their publications, where the nodes represent individual researchers, and the links denote co-author relationships.

From such a bibliography network, we aim to discover "rising stars". To do so, we 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 model the degree of mutual influence using a novel link weighting strategy. 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 and conferences is more likely to be an influential researcher as compared to another who publishes at less significant venues. This is accounted for by placing different weights on different nodes in the network

model. 3) *The chronological changes in 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.

In this work, we design a novel PubRank algorithm to mine rising stars from bibliography networks which incorporates the factors described above. Our algorithm derives information from the out-links of nodes, which is fundamentally different from many related node analysis algorithms that use information from the in-links.

Our technique is potentially useful for academics 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.

We have also implemented a graphical interactive system RStar for public access to our results on the DBLP data (http://rstar.i2r.a-star.edu.sg/).

2 The Proposed Technique

Constructing the bibliography network. A bibliography network is a directed, weighted network where the nodes represent authors and the edges denote co-author relationships. 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. We model this mutual influence using the number of publications co-authored as a proxy for the strength of their collaboration relationship. We set the weight of the edge (v_i , v_j) to be the fraction of author $\boldsymbol{v_j}$'s publications that were co-authored with author $\boldsymbol{v_i}$, and the weight of the edge (v_j, v_i) to be the fraction of author $\boldsymbol{v_i}$'s publications that were co-authored with author $\boldsymbol{v_j}$. Researchers are then modeled to influence each other according to the strength of this relationship. Our weighting scheme captures the intuition that an expert researcher will tend to influence a junior researcher more than the junior influences the expert, as the expert will tend to have more publications, thus reducing the fraction of co-authored work with the junior researcher.

Accounting for the quality of publications: assigning node weights. The reputation and impact of a researcher is decided by the quality of his/her work. We incorporate this information by assigning node weights using the quality of a researcher's publications. While the citation count of a paper is commonly used as a measure of its quality, it is biased towards earlier publications because articles need time to accumulate citations. Rising stars, being junior researchers, are thus unlikely to have many highly cited papers. We therefore opted for an alternative measure based on the prestige of its publication venue. Numerous ranking schemas are available for this purpose. A commonly used system is as follows: rank 1 (premium), rank 2 (leading), rank 3 (reputable) and unranked [4].

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

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

where pub_i is the *i*-th publication, $r(pub_i)$ is the rank of publication pub_i, and α ($0 < \alpha < 1$) is a damping factor 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.

Propagating influence in the bibliography network. 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. This feedback nature has also been famously observed in the "social network" of co-referencing web-pages, and is exploited by Google's PageRank algorithm [5], where the PageRank of a page is defined in terms of the PageRanks of pages that link to it. We adapted the PageRank algorithm to compute a similar score for each node based on the propagation of influence in the bibliography network. We account for the mutual influence between authors and the quality of each author's publications to compute a similar PubRank score for each author (node):

$$PubRank(p_i) = \frac{1-d}{N} + d * \sum_{j=1}^{|V|} \frac{w(p_i, p_j) * \lambda(p_i) * PubRank(p_j)}{\sum_{k=1}^{|V|} w(p_k, p_j) * \lambda(p_k)} \quad (2)$$

In equation 2, N is the number of authors in the network, $w(p_i, p_j)$ is the weight of the edge (p_i, p_j) and $\lambda(p_i)$ is the publication quality score defined in equation 1.

A key difference between the PageRank and PubRank scores is that the PageRank score of a node is influenced by the scores of *nodes that link to it*, while the PubRank score of a node is dependent on the *nodes to which it links to*. In other words, unlike the PageRank algorithm and 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 difference reflects the reality of the situation, as a researcher who has high quality publications and is able to contribute to the work of other influential researchers is likely to be a rising star.

Discovering rising stars from the evolving networks. The bibliography network grows larger each year as more papers are published. To account for this evolution of the network, we compute a series of PubRank scores for each author over several years. We hypothesize that if a researcher demonstrates an increase in his/her annual PubRank scores that are significantly larger than those of an average researcher, he/she will probably do very well in the coming years. We thus use a linear regression model to compute a gradient for each author, regressing his/her PubRank scores during a past time period against time (as measured in years). We then assess the significance of the gradient by computing its Z-score. Assuming that the gradients of the researchers have a Gaussian distribution, a critical region typically covers 10% of the area in the tail of the distribution curve. Thus, for a researcher v_i, if his/her Z-score is larger than 1.282, we regard v_i as statistically significant — v_i will be predicted as a rising star. In addition, we also require the researcher's PubRank score at the start of the time period to be lower than the average PubRank score of all researchers. This allows us to search for the "hidden" rising stars.

3 Experimental Evaluation

We performed two experiments using publication data from the Digital Bibliography and Library Project (DBLP). Our first experiment used all the DBLP data. This large data set with over one million publications tests the scalability of our algorithm. 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 due to its long pedigree and relevance to our work in data mining. In our experiments the damping factor α was set to 2.

Results on the entire DBLP dataset. We used data from 1990-1995 to predict the rising stars, then look at their eventual PubRank scores a decade later in 2006 to verify if they have indeed realized their predicted potentials. 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), our method identified 4,459 rising stars. We compared the rising stars with researchers in general. 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). The predicted stars have indeed increased their PubRank scores significantly faster than researchers in general. 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 was 2.92, which means that they score well above that of the average researchers.

We performed a more in-depth analysis of the citation count of the top ten predicted rising stars, comparing them to the citation counts for 100 randomly selected non-rising star researchers. We found that the rising stars obtained significantly higher citation counts for their most cited papers, obtaining 440 citations on average as compared to 18.9 citations for randomly selected researchers.

We also 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 count higher or equal to *h*, is a useful index to characterize the scientific output of a researcher [6]. 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. A list of database conferences was obtained from schema [4] and we retrieved 19474 papers published at these venues from the DBLP data. Our PubRank algorithm is then 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 showed that 43 (68.3%) have been appointed full professors at renowned universities, 7 (11.1%) of them are key appointment holders at established

research laboratories and companies, and the remaining 13 are either Associate Professors or hold important positions in industry.

Table 1. Top 10 predicted rising stars from the database domain from years 1990-1994.

Name	Position and 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, Nat. Taiwan U	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, U of Pittsburgh	IEEE fellow	171
Jiawei Han	Professor, UIUC	ACM fellow	6158

Table 1 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 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 etc). Such achievements clearly show that they have indeed become the shining stars in the database domain, as we have predicted with our algorithm with publication data from more than a decade ago.

References

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

[2] 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.

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

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

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

[6] J. E. Hirsch, "An index to quantify an individual's scientific research output," *PNAS,* 2005.