Finding Rising Stars in Social Networks

Ali Daud[1], Rashid Abbasi[1] and Faqir Muhammad[2]

[1]Department of Computer Science and Software Engineering, Sector H-10, International Islamic University, Islamabad 44000, Pakistan
[2]Department of Business Administration, Sector E-9, Air University, Islamabad 44000, Pakistan
ali.daud@iiu.edu.pk, abbasi_20150@yahoo.com, aioufsd@yahoo.com

Abstract. This paper addresses the problem of finding rising stars in academic social networks. Rising stars are the authors which have low research profile in the beginning of their career but may become prominent contributors in the future. An effort for finding rising stars named PubRank is proposed, which considers mutual influence and static ranking of conferences or journals. In this work an improvement of PubRank is proposed by considering authors' contribution based mutual influence and dynamic publication venue scores. Experimental results show that proposed enhancements are useful and better rising stars are found by our proposed methods in terms of average citations based performance evaluation. Effect of parameter alpha and damping factor is also studied in detail.

Keywords: Rising stars, author contribution, dynamic publication venue score, PageRank, Academic Social Networks

1 Introduction

Academic social networks are made up of co-author and citation based relationships between authors and research papers, respectively. Co-author means the authors writing paper together and citation based relationships occur when one paper cites other papers or is cited by other papers. Academic social network analysis has many interesting research tasks such as expert finding [7], author interest finding [6] citation recommendations [8] name disambiguation [17] and rising star finding [20]. This work is focused on finding rising stars. The motivation is to find new born researchers with abilities to become stars or experts in future. All those persons, who may not be at the top at the moment or are not experienced, but are capable to be at the top position in their respective fields in near future, are referred to as Rising Stars. Finding rising stars is very useful for appointing young faculty members to increase research productivity of department, finding reviewers for conferences and journals which can provide reviews on time and making them members of different academic committees to get benefit from their dynamic and energetic behavior.

An effort made for finding rising stars considers mutual influence and static (a ranking list of publication venues) importance named PubRank [20]. The idea was if a junior author influences/collaborates with a well known researcher and the publication venue is of high rank (1,2,3 where 1 is higher level, 2 is normal level and 3 is low level) he has bright chances to become a star in future. There were two major problems with the existing method (1) the authors did not consider the author contribution oriented mutual influence in PubRank which is very important when one wants to calculate the influence of one author on another. Here, contribution means the order in which the authors appears in the paper such as first author, second author and so on, with first author is usually considered the main contributor of that work. The junior author who influences/collaborates with well known researchers as main contributor of work has more chances than that of a junior author simply influencing/collaborating with well known researchers and (2) using static rankings is not practical as quality of work published in publication venues changes every year so as the ranking of publication venues and old static ranking lists available on the web does not provide latest rankings of publication venues. Due to aforementioned reasons we are motivated to propose StarRank algorithm which overcomes the limitations of PubRank in easy way. Our proposed method considers author contribution based mutual influence of authors on each other's in terms of order in which authors appears in the paper as well as latest (dynamic) scores of rankings for publication venues which is calculated using entropy. Our intuition to use entropy is based on the fact that the venues which are stricter in accepting papers to their areas of research are of higher level and has less entropy as compared to the venues which are not very strict in accepting papers to their areas of research and has higher entropy. Here one thing needs to be made clear that usage of entropy for scoring publication venues is workable for conferences/journals but not for workshops as they accept topic specific papers but they do not need to be of high quality because they are not finished or top level papers mostly.

Our hypothesis is supported with the detailed experimentation which shows that our proposed StarRank outperformed existing method clearly for rising star finding task. The effect of algorithm parameters is also studied in details to find their suitable values for rising star finding task.

The major contributions of this work are (1) contribution oriented co-author weight (2) entropy based dynamic publication venue score (3) unification of contribution oriented weight and dynamic publication venue score (4) and experimental evaluation of our proposed method on the real world dataset of DBLP.

The rest of the work is arranged as follows. Section 2 provides the literature review of tasks performed in academic social networks followed by the applications of page rank in these networks. Section 3 provides the existing method with the detailed approach proposed by us for finding rising stars. Section 4 provides dataset description, performance evaluation procedure, parameter settings with results and discussions in different scenarios and section 5 finally provides the concluding remarks.

Several concepts are used interchangeably in this paper such as academic social networks and co-author networks, conferences/journals and publication venues, papers, research papers and publications etc.

2 Related Work

2.1 Tasks in Academic Social Networks

Author interest finding focuses on who have interest in writing on some topics [6]. Authors based on their areas of research chose one or more topics to work on. Expert finding addresses the task of finding authors which are well known in their area of research [7]. Online publication databases like DBLP and CiteSeer provide very useful information in which names are inconsistent, which is called named entity disambiguation task. Name entity disambiguation task have two major challenges which are (1) name sharing and (2) name variation [17].

Finding Association aims at discovering the relationships between nodes or people in social networks. Email networks also provide associations between the senders and receivers in several ways [1].With the emergence and rapid explosion of social applications and media, such as instant messaging (e.g. MSN, Yahoo, Skype) and blogs (e.g., Blogger, Live Journal) finding and quantifying the social influence of actors on each other is significant [18].The research has become too much planned and sophisticated these days. There are several challenging factors need to be considered, e.g. how to find that if someone is expert of a topic either he can be good advisor or not? [19].

The competition between researchers has become very challenging these days so they want paper writing a quick process, so finding accurate citations quickly is important [8]. Community mining in co-author networks is important problem, where authors are connected to each other by co-authorships or paper citations and thus can be modeled as interaction graphs by considering semantics-based generalized topic models [9].

2.2 Applications of Page Rank

PageRank is a very useful algorithm for ranking pages or important entity finding in graphs. TextRank [15] was proposed for extracting keywords, key phrases and sentences from the documents and comparable results with supervised learning algorithms are achieved. A weighted directed model AuthorRank for co-authors network is proposed by Xiaoming et al., [13]. The importance of an individual author and its popularity is weighted through prestige. People tag resources in the web according to their understanding of those resources results in developing social tagging systems which have emerged quickly on the web. FolkRank [11] algorithm is proposed to rank users, tags and resources on the basis of undirected links between them. An important area of research in Bioinformatics is biological network analysis. Personalized PageRank [12] is proposed to find important proteins in protein networks. Finding rising star is investigated through mutual influence and static ranking of conferences/journals. Mutual influence did not consider author contributions and static rankings usage is not correct as rankings of conferences/journals keeps on changing so are dynamic, which motivated us to propose StarRank [20].

3 Finding Rising Stars

In this section, before describing our (1) author contribution based, (2) dynamic publication quality based, and (3) composite StarRank approaches, we briefly introduce related existing approach PubRank [20] for rising star finding.

3.1 PubRank

PubRank method [20] was proposed to find rinsing stars from academic social networks. It can be used to find authors which can be future experts. They considered two main points (1) Mutual influence among the researchers in term of co-authorships and (2) track record of researcher's publications in terms of publishing in different level of publication venues.

Firstly, a graph is considered in which nodes describe authors and edges describe co-author relationships for calculating mutual influence. The main idea was that if a junior researcher can collaborate with expert senior researchers or can be able to influence their work he has bright chances to be a future expert. A novel link weighting strategy is proposed. When authors (A_k,A_l) are co-author in any article to calculate the weight of author $A_k$they put the weight A_k= (A_k,A_l) as fraction of A_l author which is co-author with A_k. Moreover, the weight of A_l = (A_l ,A_k) is fraction of A_k. This weighting strategy is based on the intuition that a junior researcher will influence its senior researcher less and senior will influence more as he has more publications, which reduces the junior researcher fraction of co-authored work. The following example explains how the influence is calculated.

Suppose we have two authors K with 4 papers and L with 3 papers. If they have co-authored two papers with each other, the weight with which they influence each other is calculated as:

$$W(A_l, A_k) = \frac{(A_l,A_k)}{PA_k} = \frac{2}{4} = 0.4\,, W(A_k, A_l) = \frac{(A_k,A_l)}{PA_l} = \frac{2}{3} = 0.66 \tag{2}$$

Here, PA_l and PA_k are the total number of papers written by authors L and K. *(A_l,A_k)* is the number of co-authored papers between authors L and K. The weight *W(A_l,A_k)* with which author A_l influences author A_k is less as compared to the weight *W(A_k ,A_l)* with which author A_k influences author A_l. As the number of papers written by author A_k are 5 which are more in number as compared to A_l, So, A_k is a senior and influences A_l more.

Secondly, prestige of publication venue is considered for calculating the track record of researcher's publications. The reputation of a researchers work can by judged by number of citations his papers have which is biased towards earlier publications as publication needs time to be cited by other papers. Rising starts cannot have many highly cited papers. So the publication venue levels in which they have published the papers are considered. The main idea behind this intuition is that if a researcher is publishing in high level venues in the beginning of his career he has bright chances to be an expert in future. A static ranking scheme available on web

[14] is used with following ranks. Rank 1 (premium), Rank 2 (leading), Rank 3 (reputable) and Rank 4 (unranked). The publication quality score for each researcher is calculated by using the following equation.

$$\lambda(A_i) = \frac{1}{|p|} * \sum_{i=1}^{p} \frac{1}{\propto^{r(pub)-1}} \tag{3}$$

where, $\lambda(A_i)$ is publication quality score of author A_i all publications the larger the better,pub_i is i^{th} publication, r (pub) is publication rank of the paper and $\propto$ value is between (o<$\propto$<1) which is damping factor so low rank publications can have low scores.

Finally, mutual influence among the researchers and track record of researcher's publications is hybridized in PubRank as follows.

$$pubRank(A_i) = \frac{1-d}{n} + d * \sum_{j=1}^{|a|} \frac{W(A_j, A_i) * \lambda(A_i) * pubRank(A_j)}{\sum_{k=1}^{|a|} W(A_k, A_j) * \lambda(A_k)} \tag{4}$$

where, n is total number of scientist, $W(A_j, A_i)$ is weight for authors influencing author A_i, $\lambda(A_i)$ is publication quality score of author p_i and $pubRank(A_j)$ is the PubRank of authors linking to author A_i.

3.2 StarRank

In this section, StarRank is proposed by us to find rinsing stars. Based on the intuition of Sekercioglu [5] of quantifying coauthor contribution in a research paper we proposed author contribution based mutual influence calculation method among the researchers. We also proposed a dynamic way of calculating the scores for publication venues in comparison to simply using a static list(s), which provided out dated ranking [20]. Finally both are combined to propose Composite StarRank.

3.2.1 Author Contribution based StarRank (AC StarRank)

A graph is considered in which nodes describe authors and edges describe co-author relationships for calculating mutual influence in PubRank [20]. In addition to this information the order in which authors are appeared in the papers is also considered with first author as maximum contributor. Less contributed author score would be less and more contributed author score should be more based as discussed by Sekercioglu [5]. When a paper has more than one co-author in that case equal contribution score given to all of them is unfair. We have proposed a novel link weighting strategy based on author order based contributions. For example, a paper has 4 authors and if an author appears at number one, he will have more contribution as compared to author appeared at number four. The main idea is that if a junior researcher can collaborate with senior researchers and can influence their work as a main contributor such as by appearing at number one or two in the paper he has bright chances to become future expert.

Suppose we have four authors K, L, M, N; K with 4 papers, L with 3 papers, M with 4 and N with 3 papers with each other. The order in which they appear in the paper is given in the following. The author K and L co-authored and M and N co-authored the papers with each other highlighted as bold face letters in the following table.

Table 1. Authors with their papers and order of appearance.

Authors	Paper No (order of appearance)
K	**1(1), 2(3),** 3(2), 4(1)
L	**1(2), 2(2),** 3(1)
M	**1(1), 2(3),** 3(2), 4(1)
N	**1(3), 2(4),** 3(4)

One can see that L is junior researcher in co-authorship of K and L as L has 3 papers and K has 4 papers and N is junior researcher in co-authorship of M and N as N has 3 papers and M has 4 papers. In this co-authorship scenario L has 1 paper as first author and 2 papers as second author while N has 1 paper as third author and 2 papers as fourth author.

$$ACW(A_l, A_k) = \frac{(\sum AC_l + \sum AC_k)}{\sum PAC_k} = \frac{(0.5+0.5)+(1+0.33)}{1+0.33+0.5+1} = 0.823 \tag{5}$$

$$ACW(A_n, A_m) = \frac{(\sum AC_n + \sum AC_m)}{\sum PAC_m} = \frac{(0.33+0.25)+(1+0.33)}{1+0.33+0.5+1} = 0.67$$

Here, $\sum PAC_k$ and $\sum PAC_m$ is the total contribution of author K and M in all papers written by them. $\sum AC_l + \sum AC_k$ is the authors L and K contributions of co-authored papers. For example, author L has co-author paper 1 with author K as second author 2(2) so its contribution is ½ =0.5, which we get from 1/R where R is the rank of author [5] in the co-authored papers.

The author contribution weight $ACW(A_l, A_k)$ with which author A_l influences A_k is 0.823 which the contribution weight $ACW(A_n, A_m)$ with author A_n influences A_m is 0.67 which is less, as in the co-authored papers L is at number 2, while N is number three in one paper and number four in other paper. They both (L, N) have same number of papers and co-authored papers as well as their seniors have same number of papers but L was able to more influence his senior researcher due to more contribution in the work he has co-authored.

3.2.2 Dynamic Publication venue based StarRank (DPV StarRank)

In this section dynamic publication venue based StarRank score is calculated using entropy. We have calculated entropy of publication venue. The entropy is a measure of disorder in Physics and less disorder means the systems is better. The same phenomenon is workable here in this work as high level venues has low entropy while the non high level venues will have higher entropy. Here, it is necessary to mention that using entropy enables us to calculate venues influence for the existing and new coming venues.

In case we use existing online list of venues levels there can be different problems. Such as (1) the list may not or usually do not contains all venues and (2) new coming venues will be added later or may be missed on some occasions. Entropy of venues is calculated by using the following standard equation in which w_i is the probability of $word_i$ in a venue *v*. The title words of papers published in a venue are used for calculating the entropy of venues. The lesser the entropy is better as it corroborates that the venue has less disorder due to only accepting papers on its specific areas mentioned in call for paper page.

$$Entropy\ (v) = -\sum_{i=1}^{m} w_i log_2(w_i) \qquad (6)$$

The publication quality score for each researcher is calculated by using the following equation.

$$\lambda(dpq_i) = \frac{1}{|p|} * \sum_{i=1}^{p} \frac{1}{\propto^{\text{Entropy of Venue}}} \qquad (7)$$

where, $\lambda(dpq_i)$ is publication quality score of authorA_i all publications the larger the better,entropy of venue is publication rank of the paper and $\propto$ value is between (o<$\propto$<1) which is damping factor so low rank publications can have low scores.

3.2.3 Composite StarRank

In this section composite StarRank method is provided which calculates the rank of author according to author contribution based mutual influence and dynamic publication venue based scores.

$$StarRank(p_i) = \frac{1-d}{n} + d * \sum_{j=1}^{|a|} \frac{ACW(p_i,p_j)*\lambda(dpq)*starRank(p_j)}{\sum_{k=1}^{|v|} ACW(p_k,p_j)*\lambda(dpq)} \qquad (8)$$

where, $StarRank(p_i)$ is hybridized score for authors with the higher the score the author is considered rising star.

4 Experiments

4.1 Dataset

The dataset is taken from Digital Bibliography and Library Project DBLP [21]. The data of 1996-2000 is used to predict rising stars. The title of paper, author name and conference/journal where papers have been published are considered as data variables. Stop words are removed and lower casing is performed as standard text preprocessing steps.

4.2 Performance Measurement

The ground truth ranking of rising stars is not available. The authors ranked top using StarRank and their standings are checked later in 2012 to verify if they have realized their predicted potentials or not for the performance evaluation of our proposed methods. The number of papers and citations of papers are averaged for top ranked authors for all methods using arnetminer [2]. If an author have high number citations for his papers he is usually considered better, while high number of papers can be useful but results and discussions explains that even an author A having less number of papers than author B can have more number of citations due to the high quality of his work. PubRank [20] which determines the rising stars from academic social network is used as existing method.

4.3 Parameter Settings

The value of alpha used in our experiments is 0.5 and damping factor value is 0.85. The detailed study of selecting these values for finding rising stars is provided in the sections 4.4.3 and 4.4.4.

4.4 Results and Discussions

4.4.1 Rising Stars

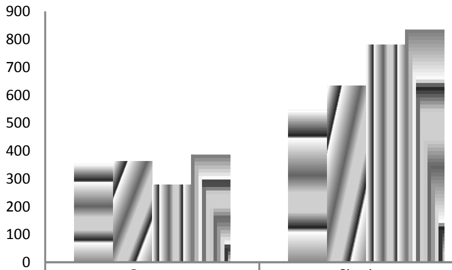
Top ten rising stars found using composite StarRank are shown in Table 2. We have visited their web pages by searching their names in Google. One can see that all of them received at least 3937 or higher number of citations for their published papers. All the top ranked authors are working on key posts now in famous IT companies as well as top ranked universities.

Table 2. Top Ten Predicted Rising Stars from StarRank.

Author	Position	Citation
Wei Ying Ma	Principal Researcher, Research Area Manager, Microsoft Research Asia	14355
Philip S. yu	Professor and Wexler Chair in Information Technology, Department of Computer Science, University of Illinois Chicago	28429
Jiawei Han	Professor, Department of Computer Science, University of Illinois at Urbana-Champaign	46654
Zheng Chen	Senior Researcher , Microsoft Research Asia	3937
Divesh Srivastava	AT&T Labs, Inc.	11520
Wei Wang	Professor, University of North Carolina at Chapel Hill	9873
Hsinchun Chen	Professor and Director, Management Information Systems Department Eller College of Management The University of Arizona	8161
Erik d. Demaine	Associate Professor, Massachusetts Inst. Tech., Lab. for Computer Science	7361
Bertram ludaumlscher	Professor of Computer Science, Computer Science Department Stanford University	29544
Lee Tan	Provost's Chair Professor, School of Computing	6824

4.4.2 Comparative Study

Top ten authors are found for all the methods. The average papers and citations of top ten authors are shown in Figure 1. For PubRank [20] the average number of papers and citations for top ten authors are 353.2 and 546.5 which are less in number as compared to our proposed three methods except for average number of papers of top ten authors for DPV StarRank which is not important. One can see that even the average number of papers for DPV StarRank is less as compared to existing PubRank [20] and one of our proposed method AC StarRank but DPV StarRank has more number of citations for top ten authors which is usually used criterion for evaluating the quality of research work published. This shows that even papers are less in number but if they are published in high level venues they are cited more which shows their popularity. It is clear from the Figure 1 that our proposed methods outperformed existing PubRank method in terms of average citation count for top ten authors.



	Papers	Citations
PubRank	353.2	546.5
AC StarRank	362	632.55
DPV StarRank	278.3	780.45
Composite StarRank	385.2	833.25

Fig. 1. Overall performance comparison

4.4.3 Effect of Alpha Parameter

Alpha is commonly used to measure the internal consistency or reliability of a psychometric test score and can take value between 0 or 1. We always observe in terms of type I errors alpha, which are always small (0.1, 0.05, .01). The smaller alpha value gets the firm proof that the alternative is correct, because the probability of type I error is reduced, but in some case high value of alpha causes high variance [16]. We calculate the rank of author using the different values of alpha 0.1, 0.2, 0.3, and up to 0.9 shown in Figure 2. When we set the alpha value 0.2 in all methods little bit change is observed in author rank, on value 0.3 author rank score is also increased a bit. For 0.4 and 0.5 value of alpha, there is maximum number of average citations and are stable. Consequently, the value of alpha used by us for all methods is 0.5.

4.4.4 Effect of Damping Factor

Damping factor value of 0.85 is usually used for ranking pages on Web [3]. Google itself uses this value because it is easy to get refined results. High damping factor means low dampened and PageRank grows higher. The StarRank is calculated for different values of damping factor to see its effect on rising star finding in terms of average citations of top ten authors. The citations of authors gradually increases from lower to high value of damping factor and one can see from Figure 3 that for 0.7, 0.8 and 0.85 maximum average citations are gained. In this paper we also used damping factor value of 0.85 though one can use 0.7 and 0.8 too. We study the effect of damping factor on average citations only as they are most important factor to judge the quality of research. As average number of papers is not a very important factor which can be considered to judge the quality of research for an author.

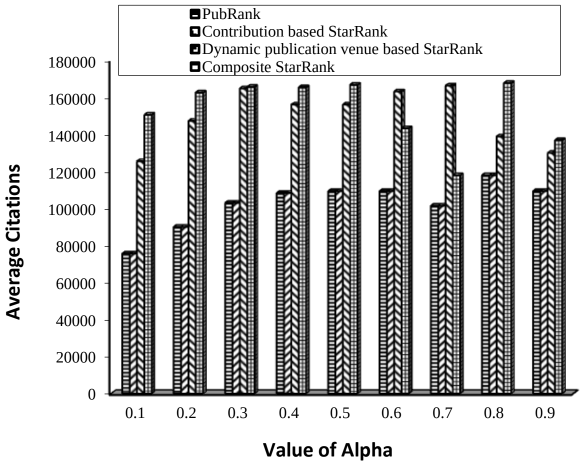


Fig. 2. Average citation on different values of alpha.

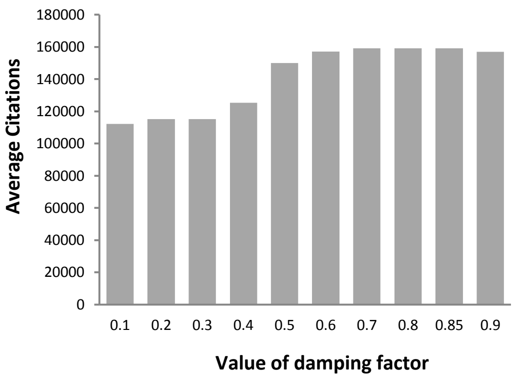


Fig. 3. Effect of damping factor in terms of Average Citations.

5. Conclusions and Future Directions

This paper addressed the problem of finding rising stars based on authors' contribution oriented mutual influence and dynamic publication venue scores. From the results we can conclude that mutual influence based author contributions are important and helped in improved results for AC StarRank method. We can also conclude that when dynamic publication venue scores are used instead of static publication venues scores, improved rising star finding results are obtained. Results have shown that dynamic publication venue scores are more useful as compared to author contribution oriented mutual influence though when both of them are hybridized more improved results are obtained. One can say that workshops are more typical venues but papers rejected from conferences are usually presented in them or incomplete works are presented in them for improvement. In future we plan to consider discriminative model for predicting rising stars as they are recently used to better predict experts for futures [10].

Acknowledgments. The work is supported by Higher Education Commission (HEC), Islamabad, Pakistan.

References

1. Adamic, L., Adar, E.: How to Search a Social Network. Social Networks, 27:187-203 (2005)
2. ArnetMiner, Academic Researcher Social Network Search, http://www.arnetminer.org/
3. Boldi, B. P., Santini, M., Sebastiano, V.: PageRank as a Function of the Damping Factor. In: Proc. of the 14th International Conference on World Wide Web, pp. 557-566 (2005)
4. Brin, S., Page, L.: The Anatomy of a Large-Scale Hyper Textual Web Search Engine. In: Proc. of International Conference on World Wide Web, pp. 107-117 (1998)
5. CH, S. Quantifying Coauthor Contributions. Science, 322 (5900):416- 417 (2008)
6. Daud, A.: Using Time Topic Modeling for Semantics-Based Dynamic Research Interest Finding. Knowledge-Based Systems (KBS), 26:154–163 (2012)
7. Daud, A., Li, J., Zhou, L., Muhammad, F.: Temporal Expert Finding through Generalized Time Topic Modeling. Knowledge-Based Systems (KBS), 23(6):615-625 (2010)
8. Daud, A., Shaikh, M. A., Rajpar, A. H.: Scientific Reference Mining using Semantic Information through Topic Modeling. Mehran University Research Journal of Engineering and Technology, 28(2):253-262 (2009)
9. Daud, A., Muhammad, F.: Group Topic Modeling for Academic Knowledge Discovery. Journal of Applied Intelligence, 36(4):876-886 (2012)
10. Fang, Y., Si, L., Mathur, A. P.: Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search. In: Proc. of SIGIR (2010)
11. Hotho, A., J¨aschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: Proc. of the 3rd European Conference on the Semantic Web: Research and Applications (ESWC), pp. 411-426 (2006)
12. Ivn, G., Grolmusz, V.: When the Web Meets the Cell: Using Personalized PageRank for Analyzing Protein Interaction Networks. Bioinformatics, 27(3):405-407 (2011)
13. Liu, X., Bollen, J, Nelson, M, L., Sompel, H. V. D.: Co-authorship Networks in the Digital Library Research Community. Information Processing and Management, 41(6):1462–1480 (2005)
14. Long, P.M., Lee, T. K. Jaffar, J.: 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)
15. Mihalcea, R., Tarau, P.: TextRank: Bringing order into Text. In: Proc. of the International Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 404-411 (2004)
16. Mohsen, T., Reg, D.: Making sense of Cronbach's alpha. International Journal of Medical Education, 2:245-246 (2011)

17. Shu, L., Long, B., Meng, W.: A Latent Topic Model for Complete Entity Resolution. In: Proc. of the International Conference on Data Engineering (ICDE) (2009)
18. Tang, J., Sun, J., Wang, C., et al.: Social Influence Analysis in Large-scale Networks. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 807-816 (2009)
19. Wang, C., Han, J., Jia, Y., Zhang, D., Yu, Y., Tang, J., J. Guo.: Mining Advisor Advisee Relationships from Research Publication Networks. In: Proc. of 16th International Conference on Knowledge Discovery and Data Mining (SIGKDD) (2010)
20. Xiao, L. L., Chuan, F. S., Kar, T. L., See-Kiong, N.: Searching for Rising Stars in Bibliography Networks. In: Proc. of the 14th International Conference on Database Systems for Advanced Applications (DASFAA), pp.288-292 (2009)
21. DBLP bibliography database, http://www.informatik.uni-trier.de/~ley/db/