Scientometrics (2013) 96:743–759
DOI 10.1007/s11192-012-0943-y

Graph-based algorithms for ranking researchers: not all swans are white!

Xiaorui Jiang · Xiaoping Sun · Hai Zhuge

Received: 10 November 2012 / Published online: 13 January 2013

Abstract Scientific importance ranking has long been an important research topic in scientometrics. Many indices based on citation counts have been proposed. In recent years, several graph-based ranking algorithms have been studied and claimed to be reasonable and effective. However, most current researches fall short of a concrete view of what these graph-based ranking algorithms bring to bibliometric analysis. In this paper, we make a comparative study of state-of-the-art graph-based algorithms using the APS (American Physical Society) dataset. We focus on ranking researchers. Some interesting findings are made. Firstly, simple citation-based indices like citation count can return surprisingly better results than many cutting-edge graph-based ranking algorithms. Secondly, how we define researcher importance may have tremendous impacts on ranking performance. Thirdly, some ranking methods which at the first glance are totally different have high rank correlations. Finally, the data of which time period are chosen for ranking greatly influence ranking performance but still remains open for further study. We also try to give explanations to a large part of the above findings. The results of this study open a third eye on the current research status of bibliometric analysis.

Keywords Scientometrics · Researcher importance · Graph-based ranking · Citation count · Recommendation intensity · American physical society

X. Jiang · X. Sun · H. Zhuge (✉)
Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No. 6 Kexueyuan South Road, Beijing 100190, China
e-mail: zhuge@ict.ac.cn

X. Jiang
e-mail: xiaoruijiang@gmail.com

X. Sun
e-mail: sunxp@kg.ict.ac.cn

Introduction

Ranking researchers is one of the most centric topics in scientometrics study. It is important for scientists to follow important researchers in their everyday research activities so as to catch up the most recent advances. It is also important for policy makers to decide who are suitable for research funding. Thus, there's a long literature history of researcher ranking. Citation count (Garfield 1972; Nerur et al. 2005) is the most simple but useful method for assessing researcher importance. Many more complicated citation-based indices have been proposed. The most famous one might be Hirsh's h-index (Hirsch 2005). Major variants of h-index include g-index (Egghe 2006), c-index (Bras-Amorós et al. 2011) and s-index (Silagadze 2010) etc.

But only using citation count based metrics for evaluation is still a controversial issue as most of these methods do not consider the network structure of literature information available. Thus, another line of graph-based methods have been developed for scientific impact evaluation (Chen et al. 2007; Walker et al. 2007; Li et al. 2008; Ding et al. 2009; Radicchi et al. 2009; Yan and Ding 2009; Zhuge and Zhang 2010). They model a literature collection as a network and apply iterative computation over the adjacency matrix of the network to achieve a converged ranking vector for objects, just as PageRank over Web pages (Brin and Page 1998). These algorithms work on one type of network. Recent works begin to integrate multiple heterogeneous networks for improve ranking (Zhou et al. 2007; Sayyadi and Getoor 2009; Das et al. 2011; Yan et al. 2011). For example, the seminal research CoRank combines the citation network and the corresponding co-authorship network and authors claim better ranking results for both authors and documents (Zhou et al. 2007).

However, there are a number of common shortcomings in current researches. Firstly, a comprehensive experimental comparison of different methods on large datasets is missing. Secondly, there is a strong need for reliable benchmark and gold standard sets. Thirdly and most importantly, there lacks some philosophical thinking about the assumptions behind these graph-based ranking algorithms and also lacks experimental support for the assumptions based one which different algorithms are designed. For example, most related researches use PageRank-like algorithms on different types of networks, such as researcher influence network, co-authorship network or author co-citation network. But which choice is better and are there any empirical evidence supporting their choice? Another example is that is iterative algorithm on researcher network a reasonable choice for ranking researchers? What is the appropriate way of defining researcher importance and is there any experimental support for the definition? As the results returned by graph-based ranking algorithms are difficult to explain and hard for information scientists to understand, it is worth studying the above questions.

In this paper, we focus on ranking researchers. We make a comprehensive empirical study of state-of-the-art graph-based ranking algorithms. We also propose to compare two simple competitive methods, using total citation count and the sum of paper ranks as researcher importance. Our findings are surprising. Simple citation count based methods beat most complicated graph-based ranking algorithms. The other method we propose, sum of paper ranks, also surprisingly becomes one of the best-performing methods. This leads us to think about the underlying assumptions of graph-based methods and above-mentioned questions. We also study what influence time factors have on ranking performances of different methods.

Methods

Datasets

In the experiments, we use the dataset provided by American Physical Society (APS). The dataset contains all the papers published by APS until 2010, the citation relationships between papers, and other metadata like authors, venues and publication dates of papers. In this study we will mostly use a subset of the whole APS dataset from 1950 to 2010. In this period, there are altogether 170,747 researchers authoring 352,882 papers. The citation network has 2,727,607 citations and its density is 7.73.

Data model

We denote the set of papers as P and the set of researchers as R. The citation network is modeled as a graph $G_C = (P, E_C)$ where each edge (i, j) denotes a citation from a paper p_i to paper p_j. In this paper, we take G_C as an un-weighted directed graph. Or equivalently speaking, we set the weights on all the edges in a citation network to 1. The authorship relationships are modeled as a graph $G_A = (P \cup R, E_A)$ where for each paper p_i and any author r_j of this paper there are two edges (i, j) and (j, i) in E_A. G_A is a typically an un-weighted un-directed graph. Based on the citation network and the authorship relationships, another two networks about researchers can be constructed: the researcher influence network and the researcher collaboration network (or equivalently co-authorship network). The researcher influence network is a graph $G_R = (R, E_R)$ where E_R is built using the citation relationships as follows. If a researcher r_i has a paper that cites another paper authored by r_j, we add an edge (i, j) to E_R. G_R might be either un-weighted or weighted directed graph. If we take G_R as weighted, the weight on edges in E_R is assigned as follows. Let

$$Cite(p_k, p_l) = \begin{cases} 1 & \text{paper } p_k \text{ cites } p_l \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

be an indicator function and $P(r_i)$ be the set of papers written by researcher r_i. The weight on edge (i, j) in E_R is as follows,

$$w_R(i,j) = \sum_{p_k \in P(r_i), p_l \in P(r_j)} Cite(p_k, p_l). \tag{2}$$

The co-authorship network is modeled as $G_{CA} = (R, E_{CA})$. for each pair of authors r_i and r_j of a paper p_k, there are two edges (i, j) and (j, i) in E_{CA}. The weight on each edge (i, j) in E_{CA} is set as the number of papers researchers r_i and r_j have collaborated on (c.f. Eq. (3)). G_{CA} is either a weighted or un-weighted un-directed graph.

$$w_{CA}(i,j) = \left| P(r_i) \cap P(r_j) \right|. \tag{3}$$

Correspondingly we can build four transition matrices. A transition matrix is a matrix of transition probabilities between states. Each row and column index is a *state* of the system. We can obtain two transition matrices for each of the above three networks as follows. (1) Let $\mathbf{P}$ be the adjacency matrix corresponding to G_C and $\widetilde{\mathbf{P}}$ is the corresponding transition matrix obtained by normalizing each row of $\mathbf{P}$. Similar to PageRank, random jump (aka. teleportation) is incorporated by rewriting the transition matrix $\widetilde{\mathbf{P}}$ into $\overline{\mathbf{P}}$ as follows:

$$\overline{\mathbf{P}} = \lambda\left(\widetilde{\mathbf{P}} + \mathbf{d}\frac{\mathbf{e}^{\mathrm{T}}}{n_{\mathrm{P}}}\right) + (1-\lambda)\frac{\mathbf{e}\mathbf{e}^{\mathrm{T}}}{n_{\mathrm{P}}} \tag{4}$$

where n_{P} is the number of papers in PIN, $\mathbf{d}$ is an n_{P}-dimensional vector indicating which row of $\widetilde{\mathbf{P}}$ is zero, $\mathbf{e}$ is an n_{P}-dimensional identity vector and $(1-\lambda)$ is the teleportation factor controlling the probability in which the rank of a paper is evenly distributed onto every paper in the network. (2) Similarly, we can construct the transition matrix $\overline{\mathbf{R}}$ of either G_R or G_{CA}. The number of researchers is n_{R}. (3) Let $\mathbf{PR}$ (resp. $\mathbf{RP}$) be the adjacency matrix from G_C (resp. either G_R or G_{CA}) to either G_R or G_{CA} (resp. G_C). We have

$$\mathbf{RP}(i,j) = \mathbf{PR}(j,i) = \begin{cases} 1 & \text{edge } (i,j) \text{ is in } G_A \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

The corresponding transition matrices $\overline{\mathbf{PR}}$ and $\overline{\mathbf{RP}}$ are constructed in a similar way. Using the four transition matrices $\overline{\mathbf{P}}$, $\overline{\mathbf{R}}$, $\mathbf{PR}$ and $\mathbf{RP}$, we put a random surfer view on ranking of papers and researchers, just as PageRank on ranking Web pages. The ranking competitors we will consider in this paper can be simply formalized in a consistent fashion.

The ranks of papers and researchers are mostly denoted as vectors $\mathbf{p}$ and $\mathbf{r}$, respectively. The only exception is that the most recently proposed "unbiased" ranking algorithm BiRank uses a randomized version of HITS-style ranking (Kleinberg 1999; Ng et al. 2001), which is a modification of the SALSA algorithm (Lempel and Moran 2001), on the citation network and thus have two paper rank vectors $\mathbf{a}$ and $\mathbf{s}$, which correspond to the *authority* and *soundness* values of papers, respectively.

Ranking competitors

Graph-based ranking algorithms

PageRank (Brin and Page 1998). There are two ways of ranking researchers by simply using PageRank algorithm. The first one is simply apply PageRank on G_R as follows

$$\mathbf{r}^{(t+1)} = \overline{\mathbf{R}}^{\mathrm{T}}\mathbf{r}^{(t)} \tag{6}$$

According to Lefebvre (2006), r will converge to the principle eigenvector of R-bar. We denote this method as PageRank in the "Results" section. The second one is ranking researchers using the ranks of papers obtained by PageRank on the citation network. Ranks of papers $\mathbf{p}$ is calculated as follows,

$$\mathbf{p}^{(t+1)} = \overline{\mathbf{P}}^{\mathrm{T}}\mathbf{p}^{(t)} \tag{7}$$

The rank of each researcher is calculated as either the sum or the average of the ranks of all his/her papers, denoted as P_SUM and P_AVG, respectively. We feel astonished that, to our knowledge, few researchers are aware of or interested in this simple method. We propose these two competitors here and our experimental results show surprisingly good results of them.

SARA (Radicchi et al. 2009). SARA applies a different weighting scheme on G_R. If paper p_i cites paper p_j and researchers r_k and r_j contribute to p_i and p_j, respectively, there is an edge (k, l) in G_R. Let the number of authors of p_i and p_j be n_i and n_j, this citation relationship will contributes $1 / (n_i \times n_j)$ to $w_R(k, l)$. SARA is originally formalized by a credit diffusion process as in Eq. (8).

$$r_i = \lambda \sum_j \frac{w_R(j,i)}{out_j} r_j + \lambda z_i \sum_j \delta(out_j) + (1-\lambda) z_i \tag{8}$$

where λ has the same meaning in Eq. (4), $out_j = \sum_k w_R(j,k)$, $\delta(out_j)$ is a unit step function of out_j and its value is 1 if out_j is bigger than zero, and z_i is defined in the following equation,

$$z_i = \frac{\sum_p Write(r_i,p) \cdot \frac{1}{|A(p)|}}{\sum_j \sum_p Write(r_j,p) \cdot \frac{1}{|A(p)|}}, \tag{9}$$

where $Write(r_i, p)$ is an indicator function indicates whether researcher r_i writes paper p and A (p) returns the set of researchers who write paper p.

SARA is in nature a variant of *personalized PageRank* and can also be formulated in matrix formalizations. After investigating Eq. (8) we know that the items $w_R(j, i)/out_j$ constitute the transition matrix $\widetilde{\mathbf{R}}$. The items $\delta(out_j)$ constitute the dangling vector $\mathbf{d}$. The vector $\mathbf{z}$ is actually the personalization vector. Thus, SARA can be rewritten into the following,

$$\mathbf{r}^{(t+1)} = \lambda(\widetilde{\mathbf{R}} + \mathbf{d}\mathbf{z}^{\mathrm{T}})^{\mathrm{T}}\mathbf{r}^{(t)} + (1-\lambda)\mathbf{z}. \tag{10}$$

CoRank (Zhou et al. 2007). CoRank is a seminal work on ranking on heterogeneous networks. In the context of literature ranking, citation network and co-authorship network are used. The ranking iteration of CoRank is modeled using both intra- and inter-network random walk. At each iteration, the surfer chooses to either walk in the current network with the probability of λ (intra-network random walk) or jump to the other network with the probability of $(1 - \lambda)$ (inter-network random walk). If the situation is the former, the random surfer will do n (resp. m) steps of intra-network random walks when he/she is currently in the citation network (resp. co-authorship network). If the situation is the latter, the random surfer will take $(2\ l + 1)$ steps of inter-network random walks. Based on the above assumptions, the equation of either paper or researcher rank has two parts and the ranks of both objects can be simultaneously obtained as in Eq. (11).

$$\begin{bmatrix} \mathbf{p}^{(t+1)} \\ \mathbf{r}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \gamma\overline{\mathbf{P}}^n & (1-\gamma)\overline{\mathbf{RP}}(\overline{\mathbf{PR}} \cdot \overline{\mathbf{RP}})^l \\ (1-\gamma)\overline{\mathbf{RP}}(\overline{\mathbf{PR}} \cdot \overline{\mathbf{RP}})^l & \gamma\overline{\mathbf{R}}^m \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{p}^{(t)} \\ \mathbf{r}^{(t)} \end{bmatrix}. \tag{11}$$

Take $\mathbf{p}^{(t+1)} = \gamma(\overline{\mathbf{P}}^n)^{\mathrm{T}} \cdot \mathbf{p}^{(t)} + (1-\gamma)(\overline{\mathbf{RP}}(\overline{\mathbf{PR}} \cdot \overline{\mathbf{RP}})^l)^{\mathrm{T}} \cdot \mathbf{r}^{(t)}$ for example, the first part depicts the intra-network ranking (in the citation network) while the second part denotes the inter-network ranking (from the co-authorship network to the citation network). The authors reported optimal results with $n = m = 2$ and $l = 1$.

FutureRank (Sayyadi and Getoor 2009). FutureRank is another state-of-the-art graph-based ranking algorithm on heterogeneous networks. The rank of a paper in FutureRank has four parts: α part from the papers who cite it; β part from its authors; γ part from its *recency* value obtained from citation dynamics analysis; and the remaining $(1 - \alpha - \beta - \gamma)$ is constantly $1/n_{\mathrm{P}}$. In FutureRank, the rank of a researcher is totally determined by the papers he/she publishes. No researcher network is used. In this paper, we omit the third part in our experiments. This does not hurt the usefulness of this study because citation dynamics is totally a different research aspect and is parallel to many other aspects like the network structure we focus in this paper and topical analysis which we do not cover here. Separating different aspects help us make clear the pros and cons of different ranking

algorithms or indices. In real applications different aspects can be integrated seamlessly. FutureRank is formalized in Eqs. (12, 13) in FutureRank

$$\mathbf{p}^{(t+1)} = \alpha\overline{\mathbf{P}}^{\mathrm{T}}\mathbf{p}^{(t)} + \beta\overline{\mathbf{R}\mathbf{P}}^{\mathrm{T}}\mathbf{r}^{(t)} + (1-\alpha-\beta)\mathbf{e}/n_{\mathrm{P}} \tag{12}$$

$$\mathbf{r}^{(t+1)} = \overline{\mathbf{P}\mathbf{R}}^{\mathrm{T}}\mathbf{p}^{(t+1)}. \tag{13}$$

P-Rank (Yan et al. 2011). In P-Rank, researcher ranks and venue ranks are used to help predict paper ranks. In this paper we will not use venue information so that all the competitors are in a fair play. The idea of P-Rank is in one way similar to FutureRank researcher importance only depends on the ranks of papers he/she writes so that no researcher network is used. But there are two differences. The first one is in P-Rank paper ranks are exclusively determined using personalized PageRank on the citation network. The second one is that researcher ranks are used to set the personalization vector $\mathbf{v}$ for all the papers. P-Rank is formalized in Eqs. (14, 15).

$$\mathbf{r}^{(t+1)} = \mathbf{P}\mathbf{R}^{\mathrm{T}}\mathbf{p}^{(t)} \tag{14}$$

$$\mathbf{p}^{(t+1)} = \mathbf{V}^{\mathrm{T}}\mathbf{p}^{(t)} \tag{15}$$

where $\mathbf{V} = \lambda\widetilde{\mathbf{P}} + (1-\lambda)\mathbf{v}\mathbf{e}^{\mathrm{T}}$ and the personalization vector $\mathbf{v}$ is set as follows,

$$\mathbf{v} = \mathbf{P}\mathbf{R}\cdot(\mathbf{r}/\mathbf{n}), \tag{16}$$

where $\mathbf{n}$ is a vector of numbers of papers for each researcher.

BiRank (Jiang et al. 2012). BiRank is the most recent multi-network ranking algorithm which aims at alleviating the problem of *ranking bias* only using the network structure. *Ranking bias* is defined as certain algorithms tending to favor papers of certain time periods and thus returning a very time-skewed top-k result list. The main ideas of BiRank are to apply a HITS-style ranking on the citation network for lowering the unreasonably high ranks of old papers and to use researcher ranks for assigning certain importance to new papers. BiRank is formulated as follows.

$$\begin{bmatrix}\mathbf{a}^{(t+1)}\\ \mathbf{s}^{(t+!)}\\ \mathbf{r}^{(t+1)}\end{bmatrix} = \begin{bmatrix}\xi\Lambda_1 & \gamma(1-\xi)\overline{\mathbf{P}}^{\mathrm{T}} & (1-\gamma)(1-\xi)\overline{\mathbf{R}\mathbf{P}}^{\mathrm{T}}\\ (1-\gamma)(1-\xi)\overline{\mathbf{P}}^{\mathrm{T}} & \xi\Lambda_1 & \gamma(1-\xi)\overline{\mathbf{R}\mathbf{P}}^{\mathrm{T}}\\ \gamma(1-\xi)\overline{\mathbf{P}\mathbf{R}}^{\mathrm{T}} & (1-\gamma)(1-\xi)\overline{\mathbf{P}\mathbf{R}}^{\mathrm{T}} & \xi\overline{\mathbf{R}}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{a}^{(t)}\\ \mathbf{s}^{(t)}\\ \mathbf{r}^{(t)}\end{bmatrix}, \tag{17}$$

where Λ_1 is a diagonal matrix with all the diagonal elements being 1, $\mathbf{a}$ and $\mathbf{s}$ are the authority and soundness vectors of papers and $\mathbf{r}$ denotes the researcher importance vector. In Eq. (17) each paper has ξ portion of its authority (resp. soundness) value directly inherited from its *most recent historical* value.

Citation-based ranking indices

We distinguish citation-based ranking indices from the graph-based ranking algorithms in that, although they both use the citation relationships, citation-based indices only need counting of number of citations without iterative calculation until convergence. In this study, we only consider the most simple and widely used citation count, both the total citation count (CC_ALL) and the average citation count (CC_AVG). We do not consider other more complex citation-based indices like h-index and g-index. The reasons are two-fold. Firstly,

h-index and g-index is typically between 0 and 200, which means there may be too many researchers having the same h-index or g-index value. This makes these indices more appropriate for a qualitative analysis. Secondly, even the results of mere citation count are good enough to reveal what we want to show. Citation-based indices are also much easier to explain and more intuitive for understanding. This is extremely appealing to information scientists, librarians and policy makers who are not computer geeks.

Evaluation metrics

There are two flavors of evaluation in the literature history. One is qualitative analysis, which is mostly applied by works in the scientometrics and information science. On the contrary, quantitative analysis is mostly favored by computer scientists. To tell clearly and intuitively, which method is better and to what extent this method outperforms its counterparts, we do a more quantitative favor of evaluation in most cases. Then, two questions remain to be answered: what is the benchmark and what is the evaluation metric? Most current works, even in the computer science domain, have not answered these two questions very well. The optimal answers to these two questions will certainly remain open questions to the research community.

Benchmark. It is very difficult for evaluating performances of different algorithms and indices on ranking papers because of lacking benchmark. It is sometimes even impossible to build such a benchmark. However, there exist some ways to build benchmarks for researchers. Following SARA, we build two benchmark sets. We collect winners of nine major Physics awards between 1960 and 2010, including Boltzmann Medal, Dannie-Heinemann Prize, Fermi Award, Gustav-Hertz Prize, Matteucci-Medal, Max-Planck-Medal, Nobel Prize, Sakurai Prize, and Wolf Prize. We do not do person name disambiguation because this task needs a lot more information which is unfortunately unavailable. So, there are some names in the dataset which may point to different persons. We do a simple treatment here. We abbreviate each author's full name using his/her last name and the first character of his/her first name. This certainly will introduce inaccuracies. For example, "Minoru Kinoshita", "Motoyasu Kinoshita" and "Makoto Kinoshita" will all be abbreviated into "M. Kinoshita". But we have two reasons to do so. Firstly, if not so we are unable to assign importance values to researchers because each author may have several names appear in the dataset. For example, the 1997 Nobel Prize winner "Philip Warren Anderson" has two alternative names in the APS dataset, "Philip W. Anderson" and "P. W. Anderson". We do not know whether we should add the values of both to assign an aggregated importance value to "Philip Warren Anderson". Secondly, when we build the benchmark, we exclude those prize winners who may have name conflicts with other researchers. The first benchmark *BenchR1* contains all the winners with at least one prize. The second benchmark *BenchR2* only contains Nobel Laureates.

Scoring. The intuition behind our scoring scheme is that, when comparing two different ranking algorithms or indices A1 and A2 on their top-k result list, A1 is better than A2 if (1) A1 returns more top-ranked researchers matching our benchmark and (2) the matched returned researchers are at the front of the top-k list. We say that the top-k list returned by a method is *recommended* by this method and we call the to-be-used metric *recommendation intensity*, which is defined as follows. If a researcher r is recommended by a method and it matches the benchmark, this method wins a score of 1. If this matched researcher is the o_r-th one on the top-k list, this method also wins an additional score of $(k - o_r)/k$. Thus, the *recommendation intensity* of a top-k result list R returned by a ranking competitor, denoted as $RI(R)@k$, is formulated in Eq. (18).

$$RI(R)@k = \sum_{r \in R} score(r)@k, \tag{18}$$

where the score of the returned researcher r in R is defined as

$$score(r)@k = \begin{cases} 1 + (k - o_r)/k & r \in BenchR \\ 0 & r \notin BenchR \end{cases}. \tag{19}$$

Recommendation intensity is in one way similar to *precision-at-k* in the information retrieval context. If we take the top-k list R as un-ordered and divide $RI(R)@k$ by k, recommendation intensity will degenerate to *precision-at-k*.

Results

Figure 1 shows the recommendation intensity curves of all the competitors discussed on *BenchR1* and *BenchR2,* respectively. Parameter tuning is an intrinsic difficult problem for all the graph-based ranking algorithms. No single optimal parameter setting exists for ranking on different datasets, with different networks and weighting schemes. The parameter settings we use in this paper are as follows. λ is set to 0.85 for all the methods. For CoRank $\gamma = 0.5$, $m = n = 2$, $l = 1$. For FutureRank $\alpha = 0.5$, $\beta = 0.4$. For BiRank, $\xi = \gamma = 0.5$. These parameter settings are proved to return rather good results experimentally in their original papers.

From both Fig. 1a, b, P_SUM, CC_ALL and FutureRank are consistently the three most effective methods. BiRank and CoRank are the next two most promising methods. What astonishes us is that simple counting-based indices perform surprisingly well. From the figures we see that CC_AVG is even much better than PageRank and SARA, not to speak CC_ALL which beats most complicated graph-based ranking algorithms. Figure 1 gives us a broad picture and leave us several problems to dig deeper.

1. Is this picture still true for different time periods?
2. What are the common and uncommon points that lead to the similarities and dissimilarities between the performance curves of different methods?
3. Why are certain methods better than other competitors?
4. What are the implications of these results to us?

In the following we try to give answers to a large part of these questions.

Time periods

For clarity, in the following we will only represent results of P_SUM, CC_ALL, FutureRank, CoRank, SARA and P-Rank in most cases. The reasons are the following. (1) The first three have been identified as the three most effective ones. (2) CoRank is slightly better than BiRank in ranking researchers and is chosen as a representative. (3) SARA is a recent method designed for researcher impact ranking. (4) P-Rank is the most recent work from the domain of information science and it bears similarities to some other methods in many ways.

Figure 2 illustrates the performances of different methods in four different time periods. We see that P_SUM, CC_ALL and FutureRank are still the three best methods. One little exception happens in Fig. 2d where the recommendation intensity curves of CoRank and CC_ALL are twisted. An interesting phenomenon is that positions of P_SUM and CoRank

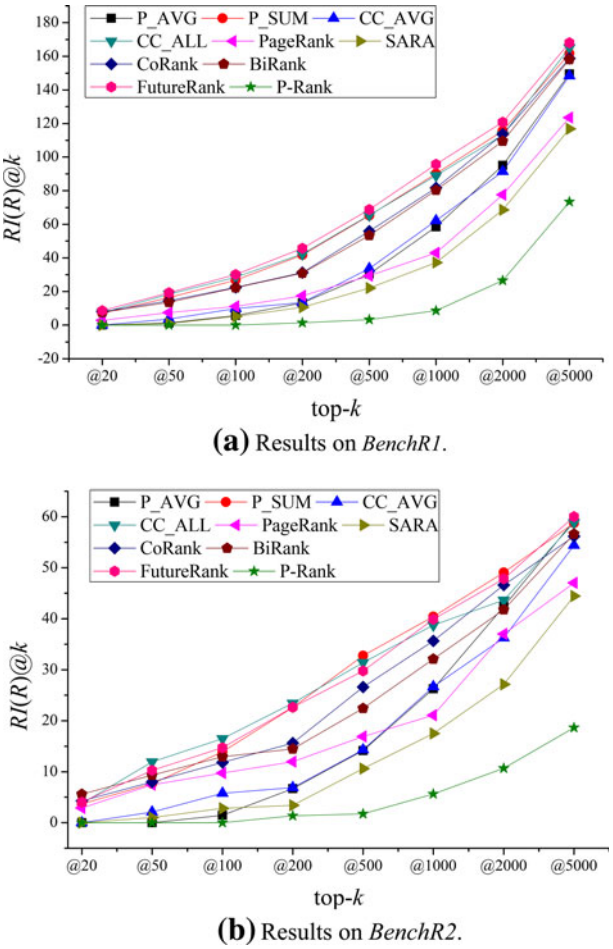

Fig. 1 Results on data between 1950 and 1970

among the six competitors are stable. CC_ALL drops down when newer data are incorporated. Graph-based ranking algorithms are more stable. To understand this phenomenon better, we do the following experiment to see what happens when more new data are added for researcher importance evaluation. We select four time periods with the start year fixed to 1950 and the end year ranging from 1970 to 2010. We test P_SUM, CC_ALL, CoRank and FutureRank on the data within each time period to see how the performance changes. Figure 3 shows the results.

The tendency shown in Fig. 2 can be seen more clearly in Fig. 3. When newer data are added, performances of P_SUM, CoRank and CC_ALL first climb and then go down. The difference lies in the time and degree of this turn. We see CC_ALL drops sharply when publications after 1980 are incorporated. While CoRank has the same tendency but its curve is flatter. The explanation of this grow-and-drop phenomenon is as follows. From one hand, the benchmark set we use consist of all the big award winners from 1960 to 2010. Many of these famous scientists are born in the first half of this century and most of their important work leading them to big awards might be done during the 1950s to 1970s.

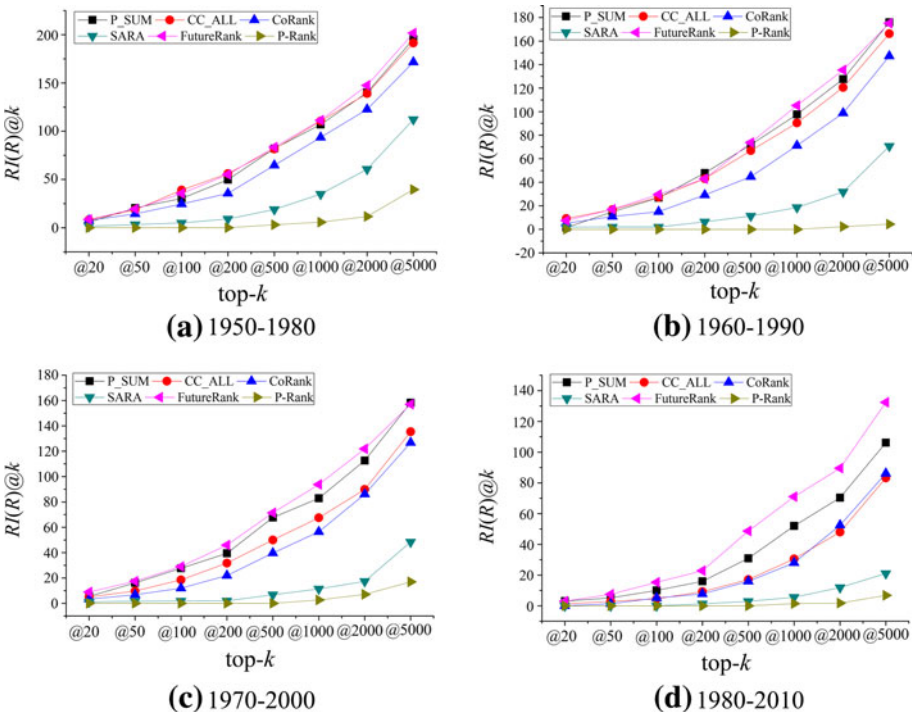

Fig. 2 Ranking results of six methods on *BenchR1* using data of different time periods with the same length

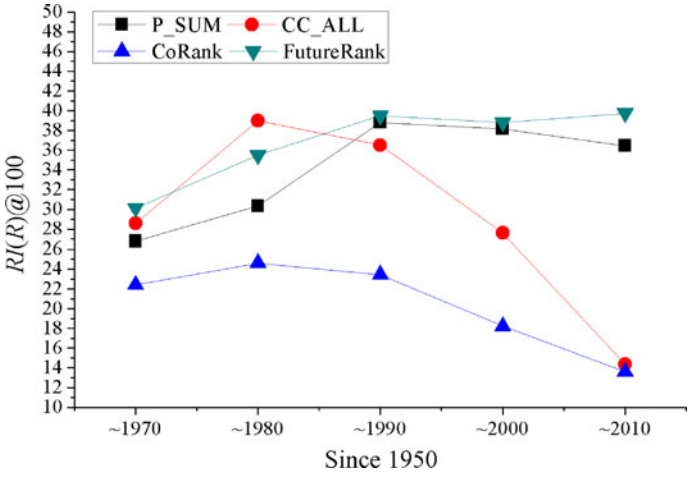

Fig. 3 Ranking results of the four best methods on *BenchR1* using data of different time periods

From the other hand, there exists a "recency" phenomenon in citation dynamics that researchers tend to cite recent publications more often than old classics (Hajra and Sen 2004; Hajra and Sen 2006; Sayyadi and Getoor 2009; Wang et al. 2009; Eom and

Fortunato 2011). Therefore, if we use relatively early data even citation counts can predict the importance Nobel Prize laureates very well. However, when more and more recent data are added, more and more younger scientists get opportunities to gain citations and accumulate personal prestige. Some of them even overtake these famous awardees even though the latter also accumulate more citations.

This is possible. Let's see Table 1 where citations of some top-ranked Nobel Prize laureates. We see that citations of Nobel Prize laureates do slightly increase from 2000 to 2010. However, some other researchers have won much more citations during this decade and, as a result, the ranks of Nobel Prize laureates using CC_ALL even decrease a little. Take the 1972 Nobel Prize laureate "J. Shrieffer" for example. "J. Perdew", "G. Gresse" and "S. Das Sarma" both have big promotion of their absolute ranks during 2000–2010, while, "D. Scalapino" and "R. Birgeneau" get relatively higher rank than "J. Shrieffer" although their positions on the top-*k* list drop a little. We have to note that some researchers win a big boost in their citations and ranks because they have name conflicts with others and the citations of different researchers are accounted to one "virtual" scientist. This happens mostly to Asian researchers many of whom have name conflicts with a few other scientists. "S. Lee" (ranked 14th in 2010, typically a Chinese Taipei or Hong Kong name), "Y. Wang (ranked 15th in 2010, typically a Chinese Mainland name)", "Y. Tokura" (ranked 20th in 2010, typically a Japanese name) and "H. Kim" (ranked 21th in 2010, typically a Korean name) are four such examples. We have to introduce more sophisticated name disambiguation algorithms which inevitably need more information that the current APS dataset cannot provide.

This problem is less severe for CoRank and P_SUM because these two methods using paper ranks to define researcher importance (see the next subsection for details). Because of the "recency" phenomenon, new papers tend to gain many citations in recent years and their ranks increase. These papers may in turn transfer a large portion of their ranks to the old classics they cite. This rank flow increases the ranks of old publications and as researcher

Table 1 Top-ranked Nobel Prize laureates

Laureate	*1950-2010*		*1950-2000*		*Year of Nobel*	*Year of Birth*
	Rank	*#cite*	*Rank*	*#cite*		
Schwinger; J.	100	3163	31	2684	1965	1912
Bethe; H.	281	2123	99	1852	1967	1908
Cooper; L.	393	1863	169	1444	1972	1930
Schrieffer; J.	**35**	4609	**17**	3505	1972	1931
Glashow; S.	219	2384	93	1905	1979	1932
Weinberg; S.	26	5274	15	3966	1979	1933
Bloembergen; N.	506	1733	164	1450	1981	1920
Wilczek; F.	285	2114	205	1354	2004	*1951*
Glauber; R.	141	2813	63	2180	2005	1925
Other Top-Ranked Physics Scientist	*1950-2010*		*1950-2000*		*#cite since 2000*	*Year of Birth*
	Rank	*#cite*	*Rank*	*#cite*		
Perdew; J.	**3**	8097	**19**	3363	4734	1943
Kresse; G.	**18**	6172	**1589**	492	5680	1967
Das Sarma; S.	**27**	5023	**59**	2227	2796	1953
Scalapino; D.	30	4859	23	3103	1756	N/A
Birgeneau; R.	34	4656	28	2783	1873	1945*

importance is defined using paper ranks, these researchers also gain more prestige. We do not see grow-and-drop phenomenon in FutureRank from Fig. 3. This demonstrates the superiority of FutureRank. However, we believe that performance decrease of FutureRank might also be unavoidable when more data are added. We will discuss this more in the following two subsections. The above discussions raise a very important question.

What dataset is the most appropriate for ranking?

For a designed benchmark or gold standard set, it is not necessarily the huge dataset which is the most appropriate. Time factors studied above should attract more research when designing test benchmarks.

Citation network versus researcher network

We have already seen that P_SUM is one of the best performers and returns comparatively good results to FutureRank. The fact that P_SUM defines researcher importance as the sum of the ranks of his/her papers and paper ranks are totally determined using citation network raises a second important question.

What is the right way to define or estimate researcher importance?

Taking the fact that CC_ALL is also competing, the statement becomes equivalently "whether we should use citation information between papers or use the derived researcher network to approximate ranks of researchers".

The experimental results imply that defining researcher importance in terms of his/her paper ranks is better. Let's consider an imaginary situation illustrated in Fig. 4. Suppose researcher r_0 has only 10 papers, each been cited once. Each of the 10 citing paper has a co-author who is famous with some other less famous co-authors, for example their students. We denote these famous citing researchers r_1, …, r_{10}. Now we will see how r_0 "steals" or "absorbs" an inappropriate portion of importance from the famous researchers. If all the citing papers $p_1^*, \ldots, p_{10}^*$ are not that influential, using the citation network only, all the papers p_1,…, p_{10} of researcher r_0 will be also less influential and thus r_0 is ranked low using P_SUM (the circle around r_0 is small in Fig. 4a). However, as r_0 is cited by ten famous researchers, using the researcher influence network, r_0 has high probability of being ranked high because r_1,…, r_{10} all devote a portion of their (huge) credits to r_0 (the

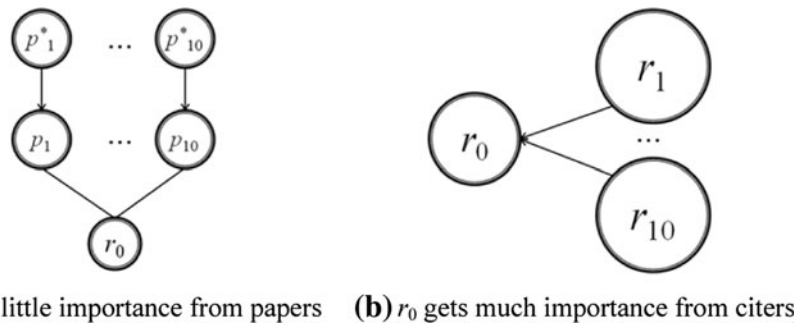

Fig. 4 Citation network versus researcher influence network on researcher ranking

circle around r_0 is relatively large in Fig. 4b). However, the reality is r_0 has only 10 publications and what's worse he/she has only gained 10 citations.

The above analysis might explain the results that P_SUM and even CC_ALL perform much better than most graph-based algorithms on researcher networks such as PageRank and SARA, as well as CoRank. This might also explain why FutureRank is also better than CoRank and BiRank, where either research influence network or co-authorship network is used for researcher ranking.

Interaction between paper rank and researcher importance

The general idea behind all the ranking algorithms using heterogeneous networks is that ranking of one type of objects has positive impacts on the ranking of other related types of objects. Take literature ranking for example, CoRank assumes that researcher ranking has impacts on paper ranking. If a paper is written by a famous scientist, it is regarded as (possibly) more influential than papers whose authors are less-known. This might be true in reality and to some extent conforms to human behavior. Many of us may routinely check the homepages of famous scientists we are interested into catch up the recent progress of his/her group. But, one thing remains open. The mutual reinforcement between ranks of papers and researcher importance is somewhat like the "chicken or egg" causality dilemma. If we want the reinforced paper rank to be correct, we must be sure that the researcher importance is appropriate, which in turn relies on the correctness of paper rank. Thus, what if the initial researcher importance is wrong as is often the case at the beginning of any iterative ranking algorithm? Is it guaranteed that the converged values of both paper rank and researcher importance are correct? Unfortunately there's no theoretical answer until now.

Our experiments show that P_SUM performs much better than CoRank and BiRank. Does it imply that paper ranks returned by PageRank on citation network is more accurate than CoRank or BiRank. Unfortunately many other studies show that this is not true. The results that P_SUM is better do not imply that P_SUM gives more accurate paper ranking. On the contrary, its paper ranking might be worse than CoRank and BiRank. So we guess that it is *the way researcher importance is defined* that finally affects the performance of a method. This is just what we have discussed in the previous subsection. The reason that CoRank and BiRank are worse might be that their calculation of researcher importance also relies on the researcher network, which has already been proved by example to be error-prone. Another evidence of this conjecture is from FutureRank. FutureRank is also a multi-network ranking algorithm but performs rather well. The reason might be the *positive feedback* between paper rank and researcher importance. In FutureRank, researcher importance is totally relied on paper rank. As we have just shown, in this way, researcher importance in FutureRank might be less erroneous than CoRank or BiRank. Then when ranking papers, researcher importance might impose a positive impact on paper rank which in turns positively influences the researcher importance in a new iteration. The main cause of this positive feedback might be also be the way FutureaRank defines researcher importance.

Correlation analysis

Correlation analysis is extensively used by information scientists to see how much new information is provided by one method compared to another. In Table 2, we list the top-ranked Nobel Prize laureates who are ranked as the top-200 by at least one ranking method and their ranks by P_SUM, CC_ALL, CoRank, SARA and FutureRank (c.f. the O_{abs}

Table 2 Top-ranked Nobel Prize laureates using 1950–1980 data

Laureates	P_SUM		CC_ALL		FutureRank		CoRank		SARA	
	O_{abs}	O_{rel}	O_{abs}	O_{rel}	O_{abs}	O_{rel}	O_{abs}	O_{rel}	O_{abs}	O_{rel}
Hofstadter; R.	38	5	55	7	42	5	80	6	822	5
Wigner; E.	293	16	134	11	360	16	402	13	5,523	17
Townes; C.	47	8	161	12	272	13	140	8	1,792	9
Schwinger; J.	11	1	1	1	13	2	4	1	707	4
Bethe; H.	24	3	26	4	19	3	23	4	878	6
Cooper; L.	45	7	65	9	56	8	315	11	4,146	14
Schrieffer; J.	44	6	63	8	30	4	168	9	1,843	10
Bohr; A.	144	14	167	14	173	12	791	16	4,888	15
Glashow; S.	126	13	105	10	43	6	226	10	2,321	12
Weinberg; S.	36	4	12	2	9	1	12	2	124	1
Bloembergen; N.	49	9	48	6	65	10	45	5	188	3
Lederman; L.	176	15	506	16	303	15	491	14	1,042	7
Ramsey; N.	20	2	20	3	52	6	15	3	173	2
Brockhouse; B.	117	11	166	13	128	11	321	12	2,127	11
Shull; C.	120	12	228	16	542	17	585	15	3,949	13
Wilczek; F.	873	17	709	17	182	13	1633	17	5,325	16
Glauber; R.	110	10	39	5	59	9	94	7	1,533	8

Table 3 Spearman's coefficients of the relative ranks of top-ranked Nobel Prize laureates in Table 1

	P_SUM	CC_ALL	FutureRank	CoRank	SARA
P_SUM		0.8548	0.8150	0.8897	0.7525
CC_ALL	0.8548		0.8401	0.9528	0.7555
FutureRank	0.8150	0.8401		0.7978	0.6299
CoRank	0.8897	0.9528	0.7978		0.8676
SARA	0.7525	0.7555	0.6299	0.8676	

Table 4 Spearman's coefficients using 1950–1980 data

	P_SUM	CC_ALL	FutureRank	CoRank	SARA
P_SUM		0.9370	0.9274	0.7982	0.6240
CC_ALL	0.9370		0.8612	0.7073	0.5622
FutureRank	0.9274	0.8612		0.7237	0.6021
CoRank	0.7982	0.7073	0.7237		0.6771
SARA	0.6240	0.5622	0.6021	0.6771	

columns). Then we calculate their relative rank (c.f. the O_{rel} columns) using their absolute ranks. For example, "R. Hofstadter" is ranked 38-th by P_SUM and only four other Nobel Prize laureates "J. Schwinge", "N. Ramsey", "H. Bethe" and "S. Weinberg" are ranked higher by P_SUM, so the relative rank of "R. Hofstadter" by P_SUM is 5 while the relative ranks of "J. Schwinger", "N. Ramsey", "H. Bethe" and "S. Weinberg" by P_SUM are 1, 2,

3, and 4, respectively. Table 3 shows the Spearman's coefficients between the relative ranks of different methods. Amongst all the five methods, SARA is the least correlated to the other four methods, especially FutureRank. P_SUM and CC_ALL both have high correlation with the other two methods except SARA. FutureRank is more correlated with P_SUM and CC_ALL. These results conform to the findings in Figs. 1, 2a.

A more accurate and comprehensive way of looking at ranking correlation is to use all the data. Table 4 lists pair-wise Spearman's rank correlation coefficients using the 1950–1980 data. The results are mostly consistent with Table 3. Again we see SARA is less correlated with the other four methods. P_SUM, CC_ALL and FutureRank are highly correlated with each other (c.f Fig. 2a).

Open questions

There are a series of open questions waiting for cleverer answers from the research community.

(1) Concerning ranking researchers, the biggest question should be the way we define researcher importance. This will greatly influence our choice of method to evaluate researchers effectively. One conjecture is that researcher importance is better captured by his/her scientific products. To prove or disprove the conjecture needs further studies from social factors of the scientific community. For example it is worth studying the factors which influence decisions of the prize committees. How do they judge the importance of researchers? Do they analyze the impacts of their papers or what else? It is also very useful to study how other scientists judge the importance of a researcher. Do the humans do in a similar way to our conjecture, which is supported by empirical studies?
(2) The experimental results of CC_ALL and P_SUM suggest us that simple method sometimes may be surprisingly competitive to complex graph-based ranking algorithms. But this result depends on the dataset we use. To study the impacts of time on the choice of dataset and ranking algorithms still remains an open problem.
(3) The third problem should be the interaction between researcher importance and paper ranks, which we have shown something similar to the "Chicken and Egg" dilemma. Is there any way to do it better? Multi-network ranking algorithms were proposed aiming to tackle this problem but the results here are unsatisfactory.
(4) The last, but not the least, open question we are concerned with is the benchmark. We must weigh all the things using the same balance. Lack of agreed standard datasets and gold standards make it difficult to compare different methods. Even we have established a way of weighing researchers, just as SARA and this paper does, some other factors are still worth paying attention to. For example the time factors we study here raise the question that whether it is necessary for us to build different benchmark and gold standard sets for different datasets. Another equivalent statement is should researcher importance be estimated statically or their importance changes with time? Most current studies ignore this issue.

Conclusions

In this paper, we make a comparative study of the performances of several state-of-the-art graph-based algorithms for ranking researchers using the American Physical Society dataset. By comparing these sophisticated algorithms with some simpler methods like

citation count and sum of paper ranks, we have identified the following challenging questions worth studying. (1) How to define researcher importance greatly influence the design and effectiveness of ranking algorithms. (2) Researcher importance is a dynamic not static issue and to choose the most appropriate dataset for ranking is nontrivial. (3) The interaction between paper rank and researcher importance, which is the basis of most multi-network ranking algorithm, is so complex that there's often a "chicken or egg" dilemma which hurts its usefulness in improving ranking performance. (4) Simple citation count based methods may have much better predictive power of important researchers than most sophisticated graph-based ranking algorithms when the dynamicity of researcher importance is taken into consideration and an appropriate dataset is chosen. This paper provides a different view of the cutting-edge research of one line of scientific importance ranking and the open questions listed in the paper worth further studies.

Acknowledgments This work is supported by National Science Foundation of China (61075074 and 61070183), Natural Science Foundation of Chongqing (No.cstc2012jjB40012), and the Key Discipline Fund of National 211 Project (Southwest University: NSKD11013).

References

Bras-Amorós, M., Domingo-Ferrer, J., & Torra, V. (2011). A bibliometric index based on the collaboration distance between cited and citing authors. *Journal of Informetrics, 5*(2), 248–264.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithms. *Journal of Informetrics, 1*(1), 8–15.

Das, S., Mitra, P., & Lee Giles, C. (2011). Ranking Authors in Digital Libraries. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pp. 251–254.

Ding, Y., Yan, E., Frazho, R., & Caverlee, J. (2009). PageRank for Ranking Authors in Co-citation Networks. *Journal of the American Society of Information Science and Technology, 60*(11), 2229–2243.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics, 69*(1), 131–152.

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE, 6*(9), 1–7.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(60), 471–479.

Hajra, K. B., & Sen, P. (2004). Aging in citation networks. *Physica A: Statistical Mechanics and its Applications, 346*(1–2), 44–48.

Hajra, K. B., & Sen, P. (2006). Modelling aging characteristics in citation networks. *Physica A: Statistical Mechanics and its Applications, 368*(2), 575–582.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 4–16569.

Jiang, X., Sun, X., & Zhuge, H. (2012). Towards an effective and unbiased ranking of scientific literature through mutual reinforcements. *Proceedings of the 21st ACM Conference on Information and Knowledge Management.*

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604–632.

Lefebvre, M. (2006). *Applied stochastic processes*. New York: Springer.

Lempel, R., & Moran, S. (2001). SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Internet Technology, 19*(2), 131–169.

Li, X., Liu, B., & Yu, P. (2008). Time sensitive ranking with application to publication search. *Proceedings of the ninth IEEE International Conference on Data Mining*, pp. 893-898.

Nerur, S., Sikora, R., Mangalaraj, G., & Balijepally, V. (2005). Assessing the relative influence of journals in a citation network. *Communications of the ACM, 48*(11), 71–74.

Ng, A. Y., Zheng, A. X., & Jordan, M. I. (2001). Stable algorithms for link analysis. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 258–266.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E, 80*(5), 056103–056112.

Sayyadi, H., & Getoor, L. (2009). FutureRank: ranking scientific articles by predicting their future PageRank. *Proceedings of 2009 SIAM Conference on Data Mining*, pp. 533–544.

Silagadze, Z. (2010). Citation entropy and research impact estimation. *Acta Physica Polonica A, B41*, 2325–2333.

Walker, D., Xie, H., Yan, K–. K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics, 7*, 06010–06019.

Wang, M., Yu, G., & Yu, D. (2009). Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and its Applications, 368*(2), 575–582.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: a co-authorship network analysis. *Journal of the American Society of Information Science and Technology, 60*(10), 2107–2118.

Yan, E., Ding, Y., & Sugimoto, C. R. (2011). P-Rank: an indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society of Information Science and Technology, 62*(3), 467–477.

Zhou, D., Orshanskiy, S. A., Zha, H., & Lee Giles, C. (2007). Co-Ranking authors and documents in a heterogeneous network. *Proceedings of the Seventh IEEE International Conference on Data Mining*, pp. 739–744.

Zhuge, H., & Zhang, J. (2010). Topological centrality and its e-science applications. *Journal of the American Society of Information Science and Technology, 61*(9), 1824–1841.