Journal of Informetrics 5 (2011) 1–13

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

PrestigeRank: A new evaluation method for papers and journals

Su Cheng*, Pan YunTao, Zhen YanNing, Ma Zheng, Yuan JunPeng, Guo Hong, Yu ZhengLu, Ma CaiFeng, Wu YiShan

Research Center for Information Science Methodology, Institute of Scientific and Technical Information of China, No. 15, Fuxing Road, Beijing 100038, China

ARTICLE INFO

Article history:
Received 14 November 2009
Received in revised form 27 March 2010
Accepted 31 March 2010

Keywords:
PrestigeRank
Citation analysis
Paper evaluation
Journal evaluation
PageRank
Authority factor
Impact factor

ABSTRACT

This paper studies how missing data in the PageRank algorithm influences the result of papers ranking and proposes PrestigeRank algorithm on that basis. We make use of PrestigeRank to give the ranking of all papers in physics in the Chinese Scientific and Technology Papers and Citation Database (CSTPCD) published between 2004 and 2006. We compared PrestigeRank result with PageRank and citation ranking. We found PrestigeRank is significantly correlated with PageRank and citation counts. We also used paper citation networks to rank journals, and compared the result with that of journal citation networks. We proposed PR_{sum}, PR_{ave}, and compared both of them with citation counts and impact factor. It indicates PR_{sum}, PR_{ave} can reflect journal's authority favorably. We also discuss the advantages and disadvantages, application scope and application prospects of PrestigeRank in the evaluation of papers and journals.

1. Introduction

Citation analysis is one of the most widely used bibliometric tools for ranking papers and journals. Garfield proposed that a citation count could be used to measure the impact of publications (Garfield, 1955) and the impact factor as a tool in journal evaluation (Garfield, 1972). Because these measures are easily comprehensible and quickly obtainable, they are being used more and more widely. However, the citation counts and impact factor have intrinsic limitations (Buela-Casal, 2004; Maslov & Redner, 2008). They assume that all citations are equal, no matter whether they are from an important paper or a poor-quality paper, which is clearly unreasonable.

Many researchers refer to the Search Engine Algorithms to obtain a solution to the importance differentiation of citations. Bollen, Rodriguez, and Van de Sompel (2006) undertook a journal ranking study for journal citation data from ISI using the PageRank algorithm. Using PageRank, the SCImago Research Group defined the SCImago Journal Rank (SJR) based on the SCOPUS Database (http://www.scimagojr.com). The Journal Citation Report promulgated by ISI in 2008 used a new journal evaluation index, Eigenfactor, the calculation of which is based on the PageRank algorithm, but eliminates self-citations in journals (http://www.eigenfactor.org). Having proposed a new journal ranking algorithm based on PageRank and the HITS algorithm, Su et al. (2009a, 2009b, 2009c) made use thereof to do an empirical study of Chinese science and technology journals. Several studies (Chen, Xie, Maslov, & Redner, 2007; Li & Zhai, 2007; Luo, Yang & Ma, 2007; Ma, Guan, & Zhao, 2008; Walker, Xie, Yan, & Maslov, 2007) applied PageRank algorithm to the publication citation network for measuring the importance of scientific papers. The common factor in all these studies is the comprehensive consideration of the quantity and quality of citations to calculate the scores of journals or papers. They differentiate the importance of the citations, which

* Corresponding author. Tel.: +86 13521586238.
E-mail address: sucheng@istic.ac.cn (C. Su).

1751-1577/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2010.03.011

Table 1
Basic structure of the publication citation graph.

Citation counts	Number of articles (nodes)	Percent	Edges	Citation counts	Number of articles (nodes)	Percent	Edges
0	11,603	56.94	0	10	16	0.08	160
1	5269	25.86	5269	11	11	0.05	121
2	1779	8.73	3558	12	2	0.01	24
3	789	3.87	2367	13	8	0.04	104
4	410	2.01	1640	14	4	0.02	56
5	204	1.00	1020	15	3	0.01	45
6	125	0.61	750	16	1	0.00	16
7	74	0.36	518	17	4	0.02	68
8	51	0.25	408	18	3	0.01	54
9	21	0.10	189	26	1	0.00	26
Total					20,378	100.00	16,393

is undoubtedly more reasonable than only considering the citation counts. Nevertheless, there are at least two new questions worth considering. First, with regard to journal evaluation, is the algorithm based on a journal level citation graph able to cover the differences in citations of the paper level graph? If this is possible, we should conduct journal evaluation using a network of paper citations. Because the quality of a journal is decided by the quality of all the papers contained therein, if the evaluation of the papers can be addressed, the problem of journal evaluation will be solved. Second, in contrast to a web page link graph, there is a wide range of types of citations in papers in a publication citation graph, including traditional journal articles, as well as conference papers, books, standards, patents and network information. There is no single database that contains all these types of documents and the issue of missing documents therefore has an effect on the metric. None of the above studies on designing networks of paper citations considered how to deal with missing parts of the literature.

To solve the two issues highlighted above, we propose PrestigeRank, a new algorithm for a publication citation graph based on PageRank. We also aim to find a solution suitable for cases where there are missing papers in the database citing network.

2. Data and methodology

2.1. Data

The Chinese Scientific and Technical Papers and Citations Database (CSTPCD) is a scientific publications system developed by the Institute of Scientific and Technical Information of China. CSTPCD is based on representative domestic scientific and technical journals. It contains more than 1700 Chinese scientific and technological journals published in English and Chinese (2008) with the source journals covering mathematics, information and systems science, physics, mechanics, chemistry, astronomy, geology, biology, medicine and health, agricultural science, industrial technology, electronics and communication, computing technology, transportation, aerospace, environmental science and other disciplines. Because of its authority, the CSTPCD data are widely used in scientific and technological decision-making support, evaluation of science and technology by all levels of the national science and technology management departments, universities, research institutions, journal editorials and research workers (Pan et al., 2008).

In this study, we selected all the physics-related papers in the CSTPCD published between 2004 and 2006, including (1) papers in all 31 physics journals in the CSTPCD; (2) physics papers published in other general journals; (3) papers citing others in (1) and (2); and (4) papers between 2004 and 2006 cited by these physics journal articles. In total, these four categories of papers included 20,378 papers, which were used to build a sparse citation network matrix with 20,378 nodes and 16,393 edges (see Table 1).

2.2. PageRank

Brin and Page (1998), the inventors of PageRank, began with a simple summation equation: the PageRank of a page P_i, denoted $r(P_i)$, is the sum of the PageRanks of all pages pointing into P_i.

$$r(p_i) = \sum_{p_j \in Bp_i} \frac{r(p_j)}{|p_j|} \tag{1}$$

where B_{p_i} is the set of pages pointing into P_i, and $|P_j|$ is the number of outlinks from page P_j. Note that the PageRank of inlinking pages $r(P_j)$ in Eq. (1) is tempered by the number of recommendations made by P_j, denoted by $|P_j|$.

PageRank with a web page linking network can be expressed using adjacency matrix L:

$$L_{ij} = \left\{ \begin{array}{ll} \frac{1}{n}, & \text{if page } i \text{ links page } j, \text{ the outlink of page } i \text{ is } n \\ 0, & \text{if page } i \text{ does not link page } j \end{array} \right\} \tag{2}$$

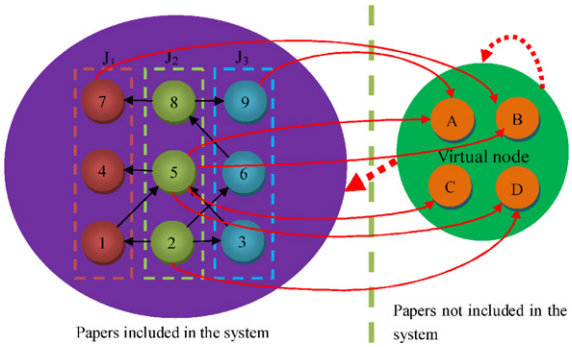

Fig. 1. Handling the missing parts in a publication citation graph.

Eq. (1) computes PageRank one page at a time. Using matrices, we replace the tedious Σ symbol, and at each iteration, compute a PageRank vector, which uses a single $1 \times n$ vector to hold the PageRank values for all pages in the index.

$$\pi^{(k+1)T} = \pi^{(k)T} L \tag{3}$$

Eq. (3) has several problems such as the problem of rank sinks, where some pages accumulate more and more PageRank at each iteration, monopolizing the scores and refusing to share. So Google's adjusted PageRank method is defined as (Langville & Meyer, 2006):

$$\pi^{(k+1)T} = \pi^{(k)T} \left(\alpha L + (\alpha a + (1-\alpha)e)\frac{1}{n}eT\right) \tag{4}$$

where $\pi^{(k)T}$ is the PageRank vector at the kth iteration, L is a very sparse hyperlink matrix, α is the scaling parameter between 0 and 1, a is the binary dangling node (pages without any out-link) vector, e^T is a row vector of all 1 s.

2.3. PrestigeRank

2.3.1. The introduction of a "virtual node"

There is a wide range of citation types in the publications citation graph. No database covers all journal papers, let alone meetings, books, network information, and other types of publications. For this study, we selected a database containing only papers, and where not all the references for the papers were included in the system.

Suppose that a system contains the publication citation graph shown in Fig. 1 with nine papers, papers 2, 5, 7, and 9 have cited papers A, B, C and D that are not included in the system, and the direction arrows between papers express the citation relationship. Solid arrow represents known citation data, dotted arrow represents unknown.

What can be deduced from Eq. (4) is that the determinants of the PageRank value of a page include the number of inlinks, the importance of other pages linking to it, and the number of outlinks of other pages linking it. So if there is a very important publication with only a small percentage of its references in the database, then the small percentage of its references will accumulate a lot of "artificial weight". As shown in Fig. 1 and Table 2, paper 5 had been cited twice and it was more important than other papers in the system, paper 5 has cited papers 4, A, B, C, D, but only paper 4 is in the system, In the context of applying the PageRank algorithm, this means that instead of distributing the weight to 5 papers, a node is distributing its weight only to one. This further means that paper 4 receives much more weight than expected.

For fixing this error we introduce "virtual node" when dealing with missing data. The "virtual node" represents those references not included in the collection and receives all citations that come from papers in the collection. In this example,

Table 2
Basic structure of the publication citation graph in Fig. 1.

	1	2	3	4	5	6	7	8	9	10	Ref[a]
1					1						1
2	1		1			1				1	4
3					1						1
4											0
5				1						4	5
6								1			1
7										1	1
8							1		1		2
9										1	1
TC	1	0	1	1	2	1	1	1	1	7	16

[a] Ref is the number of references of row of papers. TC is the times cited of column of papers, paper 10 which includes A, B, C, D is virtual node.

"Virtual node" represents papers A, B, C and D, and it is cited seven times (see Table 2 and Fig. 1). After introducing "virtual node", paper 5 cited paper 4 once and "virtual node" four times, so the weight of paper 5 which is shared by paper 4 and "virtual node" is 0.2 and 0.8, respectively. The merit of "virtual node" is that it can resolve the problem of many papers accumulating much more weight than they should, but it brings a new problem: because we know nothing about citation data of "virtual node" (dotted arrow), it is the difficult to distribute weight from virtual node to all other documents in the network.

We think the more citation counts of a paper in the system, the greater the probability of receiving citations from "virtual node". It can be expressed by a simple formula:

$$P_j = \frac{TC_j}{\sum_{i=1}^{n+1} TC_i} \tag{5}$$

where P_j is probability of the paper j being cited from outside the system, TC_j is the times cited of paper j, n is number of papers in the system, paper $n+1$ is "virtual node". The empirical observation justifying Eq. (5) complies roughly to the actual.[1]

With this method, we solve the problem of propagation of the weight from virtual node to all other documents in the network.

According to Fig. 1 and Eq. (2), if we do not introduce a virtual node, the matrix $P1$ of the publication citation graph is given by:

$$P1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

According to Eq. (5) and Table 2, if we introduce paper 10 as a virtual point, the times cited of all papers is 16, and of paper 1 is 1, the probability of being cited from paper 10 is 1/16, the times cited of paper 10 is 7, the probability of being cited from paper 10 is 7/16. So the matrix $P2$ of the publications citation graph is given by:

$$P2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 & 0 & 1/4 & 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/5 & 0 & 0 & 0 & 0 & 0 & 4/5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1/16 & 0 & 1/16 & 1/16 & 2/16 & 1/16 & 1/16 & 1/16 & 1/16 & 7/16 \end{bmatrix}$$

In matrix $P2$, rows and columns represent papers 1–10 sequentially, while the values represent the number of times that papers in the given row have cited those in the given column divided by the total number of papers cited in that row. For example, the second row of the matrix [1/4 0 1/4 0 0 1/4 0 0 0 1/4] shows that paper 2 cited papers 1, 3, 6 and virtual paper 10, and its weight is shared equally by these four papers.

[1] Firstly, we selected the papers in all journals containing "phys" in the title publishing during 2005–2008 by Chinese authors in Web of Science, and got the observed citations in this collection; secondly, we used these observed citations and Eq. (5) to get the expected citations for such papers. Finally we extended this collection across their references that are not in this collection and got new observed citations in the extended collection, we compared this expected citations with the new observed citations of these papers. We discovered the Spearman correlation coefficient between new observed and expected citations is 0.940, which suggests a correlated relationship is high. In 28% of cases the expected citations equal to new observed citation; in 63% of cases the differences between the expected citations and the new observed citations are less than 30%.

Table 3
Results of paper rankings using citation counts, PageRank and PrestigeRank.

Paper no.	Citation counts	Rank	PageRank*	Rank	PrestigeRank**	Rank
1	1	2	0.088889	6	0.068581	6
2	0	9	0.07619	9	0.053785	9
3	1	2	0.088889	6	0.068581	6
4	1	2	0.15873	2	0.075709	5
5	2	1	0.16508	1	0.13851	1
6	1	2	0.088889	6	0.068581	6
7	1	2	0.10635	4	0.085895	3
8	1	2	0.12063	3	0.096148	2
9	1	2	0.10635	4	0.085895	3
10	7				0.25831	

PageRank*: without using a virtual node; PrestigeRank**: with a virtual node; $\alpha = 0.5$.

2.3.2. PrestigeRank

After introducing "virtual node", the matrix M, of a publication citation graph containing n-papers is as follows (virtual node is $n+1$):

$$\begin{bmatrix} M_{11} & M_{12} & \dots & M_{1j} & \dots & M_{1n} & M_{1(n+1)} \\ M_{21} & M_{22} & \dots & M_{2j} & \dots & M_{2n} & M_{2(n+1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ M_{i1} & M_{i2} & \dots & M_{ij} & \dots & M_{in} & M_{i(n+1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ M_{n1} & M_{n2} & \dots & M_{nj} & \dots & M_{nn} & M_{n(n+1)} \\ M_{(n+1)1} & M_{(n+1)2} & \dots & M_{(n+1)j} & \dots & M_{(n+1)n} & M_{(n+1)(n+1)} \end{bmatrix} \tag{6}$$

$$M_{ij} = \begin{cases} \dfrac{1}{r}, & \text{if paper } i \text{ cites paper } j, \text{ the references of paper } i \text{ is } r \\ 0, & \text{if paper } i \text{ does not cite paper } j \end{cases} \quad (i \le n, j \le n)$$

$$M_{i(n+1)} = 1 - \sum_{j=1}^{n} M_{ij}$$

$$M_{(n+1)j} = \frac{TC_j}{\sum_{i=1}^{n+1} TC_i}$$

TC_j represents the times cited of paper j, n is number of papers in the system, paper $n+1$ is "virtual node".

We have defined matrix M to replace matrix L in Eq. (4) to resolve the problem of missing data in the publication citation graph. The formula for the PrestigeRank algorithm is as follows:

$$\pi^{(k+1)T} = \pi^{(k)T} \left(\alpha M + (\alpha a + (1-\alpha)e)\frac{1}{n}e^T \right) \tag{7}$$

In a web page link graph, the damping factor α is usually set to 0.85, However, Chen et al. (2007) have found that scientific papers usually follow a shorter path of about average two links and suggested the use of $\alpha = 0.5$ for publication citation graph. In this work, the authors calculate PrestigeRank values with a damping factor α of 0.50.

We selected all physics-related papers included in the CSTPCD between 2004 and 2006, a total of 20,378. According to the previous section, we can construct a 20,379 × 20,379 matrix (containing one virtual node), where the rows and columns indicate papers and values calculated by Eq. (6).

After constructing the matrix for the publication citation graph, we use Eq. (7) and a power method to seek the maximum eigenvalue of the matrix. Thereafter, we can deduce the PrestigeRank value of the 20,378 papers. Matlab was used to calculate the maximum eigenvalue of the matrix. The convergence value was set as 1E−8.

2.3.3. Evaluation of papers using different algorithms

We can calculate the PageRank and PrestigeRank value of each paper in Fig. 1 using Eq. (7), respectively, matrix $P1$, $P2$ (see Table 3). For ranking the citation counts, papers 1, 3, 6 and papers 4, 7, 8, 9 were each cited once and had the same importance. However, we examined the papers citing them carefully and found that papers 1, 3, 6 had been cited by paper 2, but paper 2 had not been cited, while papers 4, 7, 8 and 9 had been cited by papers 5, 8, 6 and 8, respectively, and since paper 5 has been cited twice and papers 6 and 8 have been cited once, they were more important than paper 2. PageRank and PrestigeRank overcome this problem. Ranking results show that papers 4, 7, 8 and 9 are more important than papers 1, 3, and 6.

Table 4
Comparison of journal evaluation results based on publication citation graph and journal citation graph.

Journal no.	PrestigeRank[1]	Rank	PrestigeRank[2]	Rank
J_1	0.230185	2	0.19391	2
J_2	0.288443	1	0.27018	1
J_3	0.223057	3	0.19391	2
J_4 (virtual node)	0.25831		0.34201	

PrestigeRank[1]: based on publication citation graph; PrestigeRank[2]: based on journal citation graph.

In Fig. 1, paper 5 was an important publication that had been cited twice, and it had 80 percent reference pointing to publications that were not included in the collection. Without virtual node, paper 4 is ranked second. With virtual node, however, paper 4 is ranked fifth. The reason for this is that without virtual node (paper 10), the weight of paper 5 is shared completely by paper 4, whereas on the contrary, the weight of paper 5 is shared equally by paper 4, A, B, C and D. Therefore, without virtual node, the importance of papers 4 increases disproportionately. The introduction of the virtual node can solve the problem of a more important paper having only a small percentage of its references in the database.

2.3.4. Comparison of journal evaluation results based on publication citation graph and those based on graph of journal citations

The importance of a journal is decided by the papers published therein, and thus the importance value of a journal based on a publication citation graph can be obtained by adding the PrestigeRank values of all the papers.

In Fig. 1 the network of paper citations contains three journals, journal J_1 containing papers 1, 4, and 7, journal J_2 containing papers 2, 5, and 8, and journal J_3 containing papers 3, 6, and 9. So the PrestigeRank value of journal J_1 is calculated by adding the PrestigeRank values of papers 1, 4, and 7; that of journal J_2 is calculated by adding the PrestigeRank values of papers 2, 5, and 8, while that of journal J_3 is calculated by adding the PrestigeRank values of papers 3, 6, and 9.

The previous study was conducted using a journal citation graph. According to Eq. (6), for the evaluation of J_1, J_2 and J_3 shown in Fig. 1, we can construct a matrix for the journal citation graph, introducing journal J_4 as a virtual node. The matrix is given by:

$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 3/11 & 0 & 3/11 & 5/11 \\ 0 & 2/3 & 0 & 1/3 \\ 3/16 & 3/16 & 3/16 & 7/16 \end{bmatrix}$$

We can ascertain the PrestigeRank value for each journal using Eq. (7) (see Table 4). As can be seen from the table, the ranking results are different when based on a publication citation graph or journal citation graph. According to the journal evaluation using the publication citation graph $J_2 > J_1 > J_3$, whereas $J_2 > J_1 = J_3$, where J_1 has the same rank as J_3, according to the journal evaluation using the journal citation graph.

Because journals J_1 and J_3 have each been cited three times, and both are cited by J_2, they are equal in the journal citation graph. However, in the publication citation graph, J_1 is cited once by papers 2, 5 and 8; while J_3 is cited twice by paper 2 and once by paper 8. From Table 3 it is clear that paper 5 is more important than paper 2, so the journals differ in that the overall citations J_1 received are more important than those J_3 received. So J_1 is better than J_3.

As can be seen from the above examples, an evaluation based on the journal citation graph covers up the citation differences of papers. On the other hand, the rank computed from the publication citation graph provides a more reasonable result for the journal evaluation.

3. Results

3.1. Results of paper rankings

3.1.1. The results of citation counts, PageRank and PrestigeRank rank

As can be seen from Table 5, there is a positive correlation in the PrestigeRank, PageRank and citation counts. The Spearman correlation coefficient between the PrestigeRank and citation count is greater than between PageRank and citation counts, which suggests a correlated relationship between PrestigeRank and citation counts is higher than between PageRank and citation counts.

As mentioned earlier, the PageRank value of one paper is in inverse proportion to the references counts of other papers citing it. But it shows a positive correlation between the percentage of references of other papers citing one paper not in the collection and PageRank of this paper (see Table 5). The results with taking no account of missing data contradict with original PageRank, because taking no account of missing data makes some papers (as paper 4 in example previously) in the collection accumulate a lot of "artificial weight". It also indicates a negative correlation between PrestigeRank and the percentage of references of other papers citing one paper not in the collection (see Table 5). Accordingly, we can draw a conclusion that introducing virtual node solves this contradiction.

Table 5
Spearman correlation coefficient of different algorithms.

Rank type	Spearman correlation coefficient
Citation counts vs. PageRank*	0.716**
Citation counts vs. PrestigeRank*	0.837**
PageRank vs. PrestigeRank	0.722**
PageRank vs. RP*	0.394**
PrestigeRank vs. RP*	−0.246**

PageRank*: without using a virtual node; PrestigeRank*: with a virtual node; $\alpha = 0.5$; RP*: the percentage of references of other papers citing one paper not in the collection.
** Correlation is significant at the 0.01 level (2-tailed).

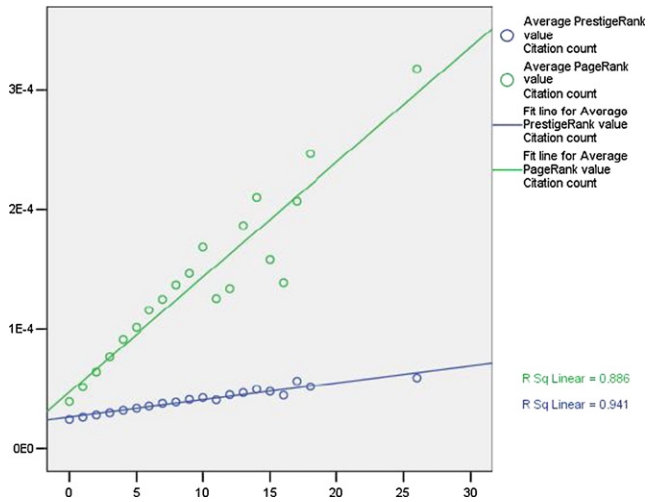

Fig. 2. Scatterplots of PreR(k), PR(k) and citation counts.

The average PrestigeRank (or PageRank) for each group of papers with K citations, namely PreR(k) (or PR(k)) is calculated instead of individual PrestigeRank (or PageRank) for each paper (see Fig. 2). Either the plot of PreR(k) or PR(k) vs. k nicely fits the line for smaller k, while the dispersion in PR(k) becomes evident when k gets larger. In Fig. 2, we can also find a correlated relationship between PreR(K) and citation counts is higher than between PR(k) and citation counts.

Table 6 gives a list of papers whose PrestigeRank, PageRank and citation counts are in the top 10. Comparing the depicted top-10 list, we observe that four papers are included in the top 10 lists for all measures (these papers are underlined in the table). We also find that the overlap between the top-10 lists produced by citation counts and PrestigeRank and citation counts is larger than the overlap between the top-10 lists obtained from PageRank and citation counts.

Most of the papers in the top 10 list are highly cited papers, but there is one modestly cited paper (paper 3104) that is highly ranked in the PrestigeRank. Tables 6 and 7 show that paper 3104 is cited three times and is ranked 4th of all papers according to the PrestigeRank and 1148th according to the PageRank, the fundamental reason for this great difference is that PrestigeRank has "virtual node" and PageRank does not have "virtual node". From further study of the papers citing them, we find that the total number of references of the 3 papers citing paper 3104 is only 8, with 7 references included in the collection, in other words, these 3 papers have a short reference list and most of their references are in the collection. Either with or without "virtual node", paper 3104 receives most of the weight of these 3 papers. Without "virtual node", many papers that are cited by papers that have a long reference list and only small part of references in the collection will receive much more weight than expected. Therefore, paper 3104 has a higher rank according to PrestigeRank than PageRank. Paper 18855 on the other hand, is just the opposite. It has a lower rank according to PrestigeRank than PageRank, Tables 6 and 8 show that paper 18855 is cited 13 times and is ranked 7th of all papers according to the PageRank and 43rd according to the PrestigeRank. From further study of the papers citing them, we can also find that the total number of references of the 13 papers citing paper 18855 is 321, with only 15 references included in the collection, in other words, these 13 papers have a long reference list and only a small percentage of their references in the collection. As mentioned above, without "virtual node", the importance of paper 18855 increases disproportionately. With "virtual node", however, paper 18855 will correctly receive only a small part of weight from these 13 papers.

In summary, based on the publication citation graph, we identified the following:

- Citation counts is significantly correlated with PageRank and PrestigeRank; PageRank is significantly correlated with PrestigeRank.

Table 6
The top 10 paper list of PrestigeRank, PageRank and citation counts[a].

Rank	Paper #	TC*	Paper #	PageRank	TC*	Paper #	PrestigeRank	TC*
1	18649	26	18649	3.18E−04	26	17016	6.487E−05	17
2	18573	18	17016	3.16E−04	17	12950	6.305E−05	17
3	18833	18	16940	3.13E−04	14	18649	5.910E−05	26
4	18951	18	18573	2.93E−04	18	19159	5.793E−05	9
5	18980	17	17266	2.79E−04	14	17266	5.445E−05	14
6	18950	17	19159	2.76E−04	9	16940	5.359E−05	14
7	12950	17	18855	2.74E−04	13	18833	5.328E−05	18
8	17016	17	19129	2.67E−04	13	18573	5.113E−05	18
9	16102	16	18833	2.54E−04	18	18951	5.074E−05	18
10	18554	15	17942	2.49E−04	11	3104	5.072E−05	3
	14309	15						
	18552	15						

TC*: times citied.
[a] Papers details:

3104: Liu WH, Spectroscopy and Spectral Analysis, 2006, 26(05), p. 865.
12950: Shao ZM, Chinese Journal of Liquid Crystals and Displays, 2005, 20(01), p. 52.
14309: Xu XJ, Chinese Physics, 2005, 14(7), p. 1287.
16102: Cai QY, CHINESE PHYSICS LETTERS, 2004, 21(04), p. 601.
16940: Xiao HR, Spectroscopy and Spectral Analysis, 2004, 24(01), p. 78.
17016: Wang DJ, Spectroscopy and Spectral Analysis, 2004, 24(04), p. 447.
17266: Yao HB, Acta Optica Sinica, 2004, 24(02), p. 158.
17942: Zhu YL, Chinese Journal of Chemical Physics, 2004, 17(02), p. 126.
18552: Zhang Y, Acta Physica Sinica, 2004, 53(02), p. 331.
18554: Hu ZT, Acta Physica Sinica, 2004, 53(02), p. 343.
18573: Li SG, Acta Physica Sinica, 2004, 53(02), p. 478.
18649: Mo JQ, Acta Physica Sinica, 2004, 53(04), p. 996.
18833: Zhao q, Acta Physica Sinica, 2004, 53(07), p. 2206.
18855: Zhao ML, Acta Physica Sinica, 2004, 53(07), p. 2357.
18950: Xie YC, Acta Physica Sinica, 2004, 53(09), p. 3020.
18951: Guo Q, Acta Physica Sinica, 2004, 53(09), p. 3025.
18980: Mo JQ, Acta Physica Sinica, 2004, 53(10), p. 3245.
19129: Wu JS, Progress In Physics, 2004, 24(01), p. 18.
19159: Li WJ, Chinese Journal of Liquid Crystals and Displays, 2004, 19(02), p. 138.

- There is a positive correlation between the percentage of references of other papers citing one paper not in the collection and PageRank of this paper and a negative correlation between PrestigeRank and the percentage of references of other papers citing one paper not in the collection. So PageRank without virtual node is unreasonable and PrestigeRank is reasonable.
- Either the plot of PreR(k) or PR(k) vs. k nicely fits the line for smaller k, while the dispersion in PR(k) becomes evident when k gets larger.
- The introduction of the virtual node can solve the problem of a more important paper having only a small percentage of its references in the database. So the result of PrestigeRank is more reasonable than PageRank.

3.2. Results of journal ranking

3.2.1. Total citation counts vs. PR_{sum} and impact factor vs. PR_{ave}

"Popularity" vs. "authority": this distinction was introduced by Kleinberg (1998) and Bollen et al. (2006), Bollen argued that the impact factor is a popularity measure of a journal, while PageRank is the authoritative measure of the journal. For journals, we can define the following two concepts: total authority PR_{sum}, authoritative factor PR_{ave}.

Table 7
Papers citing no. 3104.

Paper #	Number of references	Number of references not included in the system
3204 (Liu WH, Spectroscopy and Spectral Analysis, 2006, 26(07), p. 1264)	1	0
3344 (Liu WH, Spectroscopy and Spectral Analysis, 2006, 26(11), p. 2043)	3	1
3397 (Liu WH, Spectroscopy and Spectral Analysis, 2006, 26(12), p. 2244)	4	0
Total	8	1

Table 8
Papers citing no. 18855.

Paper #	Number of references	Number of references not included in the system
283 (Zhao ML, Chinese Physics, 2006, 15(07), p. 1611)	12	11
8643 (Zhao ML, Piezoelectrics & Acoustooptics, 2006, 28(1), p. 76)	9	8
9894 (Li XZ, Bulletin of the Chinese Ceramic Society, 2006, 25(4), p. 101)	43	42
9930 (Li XZ, Journal of Shananxi University of Technology: Natural Science Edition, 2006, 22(1), p. 8)	39	38
6508 (Xu GC, Acta Physica Sinica, 2006, 55(6), p. 3080)	20	19
9891 (Chen ZW, Journal of the Chinese Ceramic Society, 2006, 34(12), p. 1514)	53	51
9497 (Liao YW, Journal of Functional materials, 2006, 37(06), p. 886)	14	13
9599 (Liu HP, Ordnance Material Science and Engineering, 2006, 29(5), p. 52)	10	9
9602 (Chen ZW, Materials Review, 2006, 20(1), p. 14)	31	30
9637 (Su XM, Materials Review, 2006, 20(5), p. 37)	35	33
10485 (Wang XP, Materials Review, 2005, 19(10), p. 16)	24	23
15096 (Liu HP, Bulletin of the Chinese Ceramic Society, 2005, 24(3), p. 70)	20	19
15694 (Zhao ML, Journal of Functional Materials & Devices, 2004, 10(4), p. 413)	11	10
Total	321	306

Definition 1: Total authority PR_{sum}

$$PR_{sum} = \sum_{i=1}^{n} PR_i$$

Definition 2: The authority factor PR_{ave}

$$PR_{ave} = \frac{1}{n}\sum_{i=1}^{n} PR_i$$

PR_i represents the PrestigeRank value of paper i, while n represents the number of papers contained in the journal. Total authority PR_{sum} is the sum of the PrestigeRank values of the papers contained in the journal, and represents the total authority of the journal. The authority factor PR_{ave} is the PrestigeRank value of the average number of papers in the journal, and represents the mean of the authority levels of the journal papers. These two concepts correspond, respectively, to the citation counts and impact factor of the journal.

As can be seen from Table 9, we can ascertain that there is a positive correlation between PR_{sum} and the total citation counts. The Spearman correlation coefficient is 0.894 and thus they are highly related. We also find out that PR_{ave} has a positive correlation with the impact factor in 2006 (Pan et al., 2007), but the Spearman correlation coefficient is 0.417 and smaller than the correlation between PR_{sum} and the total citation counts.

Fig. 4 and Table 10 show that the total citation counts of "Chinese Physics Letters" and "High Power Laser and Particle Beams" are very different, namely 1514 and 764, but they have the same PR_{sum}, the fundamental reason for this is that PrestigeRank considers both the citation counts and quality. From Table 10 we find the total citation counts of the papers citing "Chinese Physics Letters" is 213, we also find the total citation counts of the papers citing "High Power Laser and Particle Beams" is 237, therefore, we can conclude that the papers citing "High Power Laser and Particle Beams" are more important than those citing "Chinese Physics Letters". So we think PR_{sum} is more suitable for measuring journal authority than total citation counts.

Table 9
Spearman correlation coefficient between the traditional indicator of the journal and PrestigeRank indicators.

Rank type	Spearman correlation coefficient
Total citation counts vs. PR_{sum}	0.894**
Impact factor vs. PR_{ave}	0.417*

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 10
The citation graph of papers citing "High Power Laser and Particle Beams" and "Chinese Physics Letters".

Citation counts	Papers citing High Power Laser and Particle Beams		Papers citing Chinese Physics Letters	
	Number of articles	Total	Number of articles	Total
0	620	0	1354	0
1	81	81	128	128
2	43	86	19	38
3	10	30	8	24
4	6	24	2	8
5	2	10	3	15
6	1	6	0	0
Total	763	237	1514	213

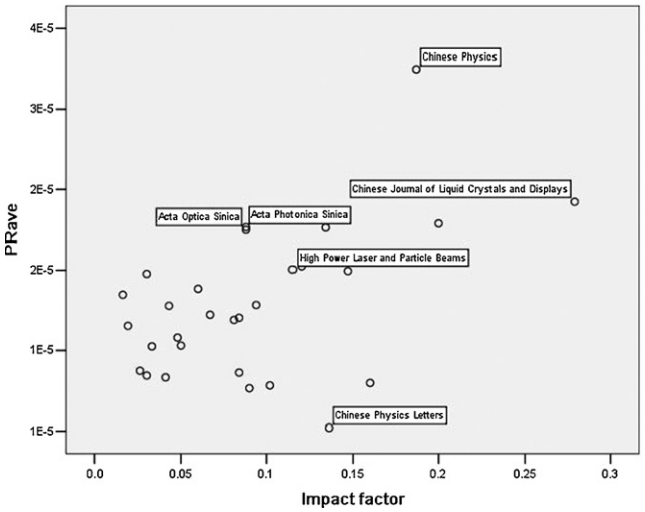

Fig. 3. Scatterplots of PR_{ave} and impact factor.

As shown in Fig. 3, "Chinese Physics Letters" has a high impact factor, but its PR_{ave} is low, "High Power Laser and Particle Beams" on the other hand, is just the opposite. As mentioned above, the main reason for this difference is that papers citing "High Power Laser and Particle Beams" are more important than those citing "Chinese Physics Letters", so we also think PR_{ave} is more suitable for measuring journal authority than impact factor.

In summary, PrestigeRank inherited from the best of PageRank, it can offset the drawbacks of traditional citation counts and impact factor that simply calculate the citation counts without taking into account the distinction of citation importance. Considering both the citation counts and quality reflects the authority of papers and journals more accurately.

3.2.2. The differences of journal ranking based on publication citation graph and those based on journal citation graph

In this work, we calculated PrestigeRank values of journals based on publication citation graph and journal citation graph, respectively (see Fig. 4 and Table 11). As can be seen from Table 11, there is a positive correlation in the PreR(p), PreR(j) and citation counts. The Spearman correlation coefficient between the PreR(p) and citation counts is 0.938, between the PreR(j) and citation counts is 0.889, and between the PreR(p) and the PreR(j) is 0.826, which suggests a correlated relationship between PreR(p) and citation counts is higher than between PreR(j) and citation counts.

Standard deviation is a widely used measure of the variability or dispersion, a low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over

Table 11
Spearman correlation coefficient of different algorithms.

Rank type	Spearman correlation coefficient
Citation counts vs. PreR(p)*	0.938**
Citation counts vs. PreR(j)*	0.889**
PreR(p) vs. PreR(j)	0.826**

PreR(p)*: PrestigeRank based on publication citation graph; PreR(j)*: PrestigeRank based on journal citation graph.

** Correlation is significant at the .01 level (2-tailed).

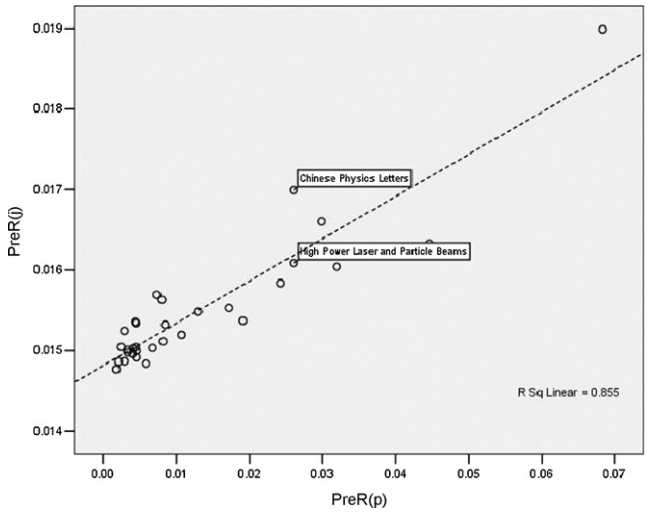

Fig. 4. Scatterplots of the PreR(p) and PreR(j) of journals.

Table 12
Descriptive statistics of PreR(p) and PreR(j).

	N	Minimum	Maximum	Sum	Range	Mean	S.D.
PreR(p)*	31	0.0018	0.0683	0.4047	0.0665	0.013055	0.0149330
PreR(j)**	31	0.014770	0.018989	0.480543	0.004219	0.01550139	0.000847056

PreR(p)*: based on publication citation graph; PreR(j)**: based on journal citation graph.

a large range of values. Range is the difference between the largest and smallest values of a numeric variable, a high range also indicates that the data are spread out over a large range of values.

Table 12 shows that PreR(p) have larger values of standard deviation and range than PreR(j), so the PrestigeRank value of journals based on graph of journal citation tends to be very close to the mean and those based on graph of paper citation are spread out over a large range of values. The fundamental reason for this phenomenon is that the PrestigeRank value based on a journal level citation graph covers the differences in citations of the paper level graph, all papers of a journal are treated equally as a whole in a journal level citation graph, which is obviously unreasonable, because the times cited of a journal cannot predict the number of citations that any individual publication will receive. Sternberg and Gordeeva (1996) points out the need to differentiate between what is published and where it is published: not everything published in the same journals is of the same quality. The correlation between the "impact" of an article and the "impact" of the journal in which it was published is far from perfect (Sternberg, 2001).

We take 2 journals (Chinese Physics Letters & High Power Laser and Particle Beams) as example in Fig. 4 below: the journal rankings based on publication citation graph are different from those based on journal citation graph. Their PreR(p) are the same if based on publication citation graph; while "Chinese Physics Letters" ranks higher than "High Power Laser and Particle Beams" if based on journal citation graph. The reason for these differences is as we discussed above: although the citation of "High Power Laser and Particle Beams" is much fewer than "Chinese Physics Letters", the importance of the papers citing "High Power Laser and Particle Beams" is generally higher than that of the papers citing "Chinese Physics Letters". Therefore, the PreR(p) of the two journals is equal. However, in the journal citation graph, all papers in a journal are deemed to be of the same importance. The ranking of journals is largely determined by the number of citations and the importance of the citing journals. The number of citations of "Chinese Physics Letters" is much higher than "High Power Laser and Particle Beams", while the higher importance of the papers citing "High Power Laser and Particle Beams" is undermined. This leads to the lower ranking of "High Power Laser and Particle Beams" at journal level. Therefore we reached the conclusion that ranking based on journal level covers the difference of the importance between the papers and therefore is definitely unreasonable.

4. Discussion and conclusion

In this work we discuss how papers that are cited by papers in the collection but are not themselves included in the collection in the PageRank algorithm influence the ranking of paper and propose PrestigeRank, which use a "virtual node" to represent those references not included in the collection and receives all citations that come from papers in the collection. We make use of PrestigeRank to give the ranking of all physics-related papers from 2004 to 2006 in CSTPCD. Furthermore, we compared the result of PrestigeRank with that of PageRank ranking and citation ranking. We also define PR_{sum} and PR_{ave} for journal ranking and compare these concepts with the impact factor and total citation counts.

It shows a positive correlation between the percentage of references of other papers citing one paper not in the collection and PageRank without virtual node of this paper, while a negative correlation with PrestigeRank. The results with taking no account of missing data contradict with original PageRank, because taking no account of missing data makes some papers in the collection accumulate a lot of "artificial weight". While PrestigeRank can solve this question by introducing virtual node.

We found PrestigeRank is significantly correlated with PageRank and citation counts. PrestigeRank inherited from the best of PageRank, it can offset the drawbacks of traditional citation counts and impact factor that simply calculate the citation counts without taking into account the distinction of citation importance. Considering both the citation counts and quality reflects the authority of papers and journals more accurately.

In this work, we compared PrestigeRank at paper level and PrestigeRank at journal level results for journal ranking with the citation ranking. We found that citation rank is highly correlated with PrestigeRank at paper level and also with PrestigeRank at journal level. We also found that the PrestigeRank value of journals at journal level tends to be very close to the mean and covers the differences in citations of the paper level graph. The results calculated with the method at paper level are more reasonable than with the PageRank algorithm at journal level.

We also found PR_{sum} has a highly correlated with total citation counts (Spearman $r = 0.894$, $p < 0.01$). This means that PrestigeRank and citation rank share similar results.

In summary, PrestigeRank presents feasible evaluation methods for papers and journals in the case where the current database system has a limited range of published papers.

This study has several limitations. (1) We have assumed that more citation counts associate with a higher probability of receiving citations from the "virtual node", but further research is needed to check this. (2) Although we found some evidence that the PrestigeRank results are more reasonable than those of PageRank in various contexts, there is insufficient evidence to make a definite conclusion that Prestige is better than PageRank or citation counts because we have not compared them against the judgements of human experts to decide which is best. Compared with citation counts and impact factors, a disadvantage of PrestigeRank is that it is more abstract and requires more calculation time.

The PrestigeRank algorithm can be used not only for the evaluation of papers, journals and other scientific entities, but also for the organization and ranking of Keywords, scientific and technical personnel, scientific institutions, and national and other scientific entities. The PrestigeRank algorithm can also be used for authoritative ranking of query results of a science and technology full-text database and has good application prospects for the organization of information.

Acknowledgments

This work is supported by the Ministry of Science and Technology in China under contract 2006BAH03B05, National Natural Science Foundation of China (70973118) and the Foundation of the Institute of Scientific and Technical Information of China (YY-200902). The authors thank three anonymous reviews for valuable comments and suggestions which helped shape and improve this paper. We also thank Xiong Ping for her assistance in polishing this article.

References

Bollen, J., Rodriguez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, *69*(3), 669–687.

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.

Buela-Casal, G. (2004). Assessing the quality of articles and scientific journals: Proposal for weighted impact factor. *Psychology in Spain*, *8*(1), 60–76.

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

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*(3159), 108–111.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Essays of An Information Scientist*, *1*, 527–544.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACMSIAM symposium on discrete algorithms* Baltimore, MD, (pp. 668–677).

Langville, A., & Meyer, C. (2006). *Google's PageRank and Beyond: The science of search engine rankings*. Princeton University Press.

Li, C., & Zhai, X. (2007). Exploration of PageRank-based citation analysis method. *Information Studies: Theory & Application*, *30*(1), 122–124.

Luo, J. Q., Yang, X. H., & Ma, J. Y. (2007). Paper auto-evaluation based on search engine. *Computer Technology and Development*, *17*(11), 80–83.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing and Management*, *44*, 800–810.

Maslov, S., & Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *Journal of Neuroscience*, *28*(44), 11103–11105.

Pan, Y. T., Ma, Z., Su, C., Guo, H., Yu, Z. L., & Xu, B. (2007). *China S&T Journal Citation Report 2007 (CORE)*. Beijing: Science and Technology Literature Press.

Pan, Y. T., Ma, Z., Su, C., Guo, H., Yu, Z. L., & Xu, B. (2008). *Chinese scientific and technical papers statistics and analysis 2006*. Beijing: Science and Technology Literature Press.

SCImago Journal and Country Rank. SCImago Research Group. Available at: http://www.scimagojr.com [Accessed: 1 October 2009].

Sternberg, R. J. (2001). Where was it published? *Observer*, *14*, 3.

Sternberg, R. J., & Gordeeva, T. (1996). The anatomy of impact: What makes an article influential? *Psychological Science*, *8*, 69–75.

Su, C., Pan, Y. T., Yuan, J. P., Guo, H., Yu, Z. L., & Hu, Z. Y. (2009). PageRank, HITS and Impact Factor for Journal Ranking. In *CSIE, 2009 WRI world congress on computer science and information engineering, vol. 6* (pp. 285–290).

Su, C., Pan, Y. T., Yuan, J. P., Ma, Z., Guo, H., Zhang, Y. H., et al. (2009b). PageRank for journal ranking. *Chinese Journal of Scientific and Technical Periodicals*, *20*(4), 614–617.
Su, C., Pan, Y. T., Yuan, J. P., Ma, Z., Guo, H., Zhang, Y. H., et al. (2009c). HITS for journal ranking. *Acta Editologica*, *21*(4), 366–369.
Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics*, P06010. doi:10.1088/1742-5468/2007/06/P06010