Scientometrics (2016) 106:805–816
DOI 10.1007/s11192-015-1805-1

Ranking scientific publications with similarity-preferential mechanism

Jianlin Zhou[1] · An Zeng[1] · Ying Fan[1] · Zengru Di[1]

Received: 21 September 2015 / Published online: 14 December 2015

Abstract Along with the advance of internet and fast updating of information, nowadays it is much easier to search and acquire scientific publications. To identify the high quality articles from the paper ocean, many ranking algorithms have been proposed. One of these methods is the famous PageRank algorithm which was originally designed to rank web pages in online systems. In this paper, we introduce a preferential mechanism to the PageRank algorithm when aggregating resource from different nodes to enhance the effect of similar nodes. The validation of the new method is performed on the data of American Physical Society journals. The results indicate that the similarity-preferential mechanism improves the performance of the PageRank algorithm in terms of ranking effectiveness, as well as robustness against malicious manipulations. Though our method is only applied to citation networks in this paper, it can be naturally used in many other real systems, such as designing search engines in the World Wide Web and revealing the leaderships in social networks.

Keywords PageRank · Citation network · Similarity · Robustness

Introduction

Following up the latest research progress is the daily task that each scientist should do. Nowadays, many new journals are created to accelerate the publication of scientific research results. Though this effort seems to make the latest findings more accessible to scientists, the reality is that the barrier for us to find the truly significant research progress

Electronic supplementary material The online version of this article (doi:10.1007/s11192-015-1805-1) contains supplementary material, which is available to authorized users.

✉ An Zeng
anzeng@bnu.edu.cn

✉ Ying Fan
yfan@bnu.edu.cn

[1] School of systems science, Beijing Normal University, Beijing 100875, People's Republic of China

becomes higher. This is due to the fact that we need to spend more and more time picking up the relevant and high quality papers from thousands of new papers. Therefore, a question raises naturally: how to objectively rank the quality of scientific publications (e.g., Radicchi et al. 2008; Maslov and Redner 2008; Frey and Rost 2010; Sorzano et al. 2014). This research problem has been studied intensively, yet the best ranking algorithm still remains to be found.

Though the Google's PageRank algorithm (Brin and Page 1998) was originally proposed to rank online web pages, it has remarkably pushed forward the research on scientific publication ranking. The PageRank algorithm provides a new perspective to the study of finding high quality articles and has broad application prospects in citation analysis. Its basic idea is that one should not only consider the number of citations that a paper received but also which papers cited it. After PageRank algorithm was introduced to scientometrics (e.g., Bollen et al. 2006; Chen et al. 2007; Fiala et al. 2008; Ma et al. 2008), a variety of modifications have been done (Fiala 2012; Nykl et al. 2014; Su et al. 2011; Walker et al. 2007; Yao et al. 2014). For instance, Su et al. (2011) propose PrestigeRank algorithm based on the results that the missing data has an impact on the rank of papers with the PageRank algorithm. Moreover, CiteRank (Walker et al. 2007) modifies the PageRank algorithm by initially distributing random surfers exponentially with age in order to account for strong ageing characteristics of citation networks. Fiala proposes a time-aware PageRank algorithm for citation networks (Fiala 2012). Yao et al. (2014) introduce non-linearity to the PageRank algorithm for further enhancing the effect of high score papers. The PageRank algorithm has also been applied to a more aggregated level such as ranking the importance of different fields, journals (González-Pereira et al. 2010), scientists (e.g., Radicchi et al. 2009; Ding et al. 2009; Yan and Ding 2011; Ding 2011), institutions (Yan 2014). An example is that the idea of PageRank is introduced to design the eigenfactor index (Bergstrom 2007; Fersht 2009) which better measures the influence of journals than the traditional impact factor (Bergstrom and West 2008).

The PageRank algorithm is an iterative process on networks. In each step, the resource of each node is updated by aggregating the resource passed from all the papers that cite it. However, this might cause some problems when PageRank is applied to real citation networks in which some noisy and malicious behaviors exist (Yao et al. 2014). For instance, some researchers might deliberately cite their own papers when they publish new papers to push up the influence of their old publications. When referring to a technical term, some authors might neglect the original literature but cite recent but less relevant publications, resulting in assigning citations to inappropriate papers. These behaviors may distort the obtained ranking of PageRank (Aksnes 2003; Foley and Della Sala 2010). In fact, many approaches have been developed to remove the noisy and spurious connections in networks (Guimerà and Sales-Pardo 2009; Zeng and Cimini 2012). The basic idea is to consider the similarity between nodes: a connection between dissimilar nodes is more likely to be a spurious connection. Inspired by this idea, we propose to redesign the resource updating rules in the PageRank algorithm according to node similarity.

In this paper, we introduce a preferential mechanism to the PageRank algorithm when aggregating resource from different nodes to enhance the effect of similar nodes. In other words, a node tends to receive resource from downstream nodes (the citing papers) that are similar to it. The method is denoted as similarity-preferential rank (SPRank). We use the citation network constructed from the data of American Physical Society (APS) journals to validate the SPRank method. We first study the basic statistics of the SPRank method and compare the ranking similarity between it and the traditional PageRank method. Then we investigate the effectiveness of SPRank. We find that SPRank can improve the ranking of

the high quality papers (e.g. Nobel prize winning papers). It can also significantly outperform PageRank in predicting the future citation growth of papers. Finally, when some papers maliciously cite a target paper to push up its citation, the SPRank method can effectively suppress the ranking of this target paper.

Method

Data collection

The database used in this paper is from the American Physical Society (APS) journals in the period from 1893 to 2009. The journals include Physical Review series, and Reviews of Modern Physics. In total, the data contains 462,720 papers with 4,620,025 citations. For each paper, the information about its DOI, authors, publication time and the DOI of its citing papers is all available in the database. Based on the above information, we construct a citation network in which nodes represent papers and edges stand for the citing relation between papers. The citation network is directed and acyclic. The indegree of a node is the number of citations it receives and the outdegree of a node is the number of papers it cites.

The SPRank method

We first briefly describe the classic PageRank algorithm. The citation network could be described by adjacency matrix A with the element $A_{ij} = 1$ if paper i cites paper j and $A_{ij} = 0$ if there is no citation relationship between paper i and paper j. We define the cited paper as the upstream node and the citing paper as the downstream node. PageRank was proposed by Page et al. for webpage ranking (1998). The algorithm can be expressed mathematically as follows:

$$s_j(t) = c + (1-c)\sum_{i=1}^{N}\left(\frac{A_{ij}}{k_i^{out}}\left(1-\delta_{k_i^{out},0}\right)+\frac{1}{N}\delta_{k_i^{out},0}\right)s_i(t-1) \quad (1)$$

where $\delta_{a,b} = 1$ when $a = b$, and $\delta_{a,b} = 0$ otherwise. In Eq. (1), N is the total numbers of nodes in the network, k_i^{out} is the number of outgoing links of node i, s_j is the PageRank score of node j through t steps iterative process and the same for s_i, c is called the return probability. The PageRank algorithm can be regarded as a random walk process on the directed network. c is the probability for a random walker to jump to a random node from the present node and $(1 - c)$ can represent the probability for a random walker to continue walking through the outgoing links of the present node. In this paper, we fix $c = 0.15$ which is the typical value of c in computer science (Brin and Page 1998). Usually the initial configuration is to set $s(0) = 1$ for all nodes. The final score of each node is defined as the steady value when $s_i(t)$ stays convergent. We define the final ranking of nodes in PageRank as R_p, which can be obtained by sorting the final scores of nodes in descending order. The notations used here for PageRank are different from those in the original paper from Brin and Page (1998) because the current notations are more convenient for introducing the following SPRank method.

The SPRank method is directly built on PageRank. The basic idea is that not all the papers in a paper's reference list is equally important to it. Normally, two papers that are topologically similar to each other must be relevant in the topic or the method. If a paper cites another paper that is similar to it, this is an "effective" citation as the citing paper

might be inspired by the cited paper or employs the algorithm in the cited paper. On the other hand, if a paper is cited by many dissimilar papers, it could be the case that the authors select references carelessly or the authors maliciously aim to push up a low quality paper. Therefore, in SPRank when the upstream node aggregates the score sent by the downstream node, the score sent by dissimilar nodes will be suppressed. In this way, even if a node is cited by many dissimilar papers, its final score cannot be high. The formula for SPRank reads as

$$s_j(t) = c + (1-c)\sum_{i=1}^{N}\left(\frac{f_{ij}^{\theta}A_{ij}}{k_i^{out}}(1-\delta_{k_i^{out},0}) + \frac{1}{N}\delta_{k_i^{out},0}\right)s_i(t-1), \tag{2}$$

where f_{ij} is the similarity between node i and j ($f_{ij} \in [0,1]$) and θ is a tunable parameter ($\theta \geq 0$). When $\theta = 0$, SPRank reduces to the classic PageRank. As θ gets larger, the score sent by dissimilar downstream nodes is suppressed more severely. We define the final ranking of nodes in SPRank as R_s which can be obtained by sorting the final score of nodes in descending order. Similarity f_{ij} in Eq. (2) could be estimated in multiple ways. In this paper, we adopt the cosine metric to measure nodes' similarity based on their outgoing links. It can be written as

$$f_{ij} = \frac{|\tau(i) \cap \tau(j)|}{\sqrt{k_i^{out}k_j^{out}}}, \tag{3}$$

where $\tau(i)$ and $\tau(j)$ are respectively the set of the upstream neighbors of node i and j. The illustration of the SPRank algorithm is shown in Fig. 1.

In fact, we also tested the case where the node similarity f_{ij} is measured by the cosine index based on nodes' incoming links. However, due to the heterogeneous indegree distribution (i.e. many nodes are with no citation and hence zero indegree), the similarity between most of the nodes is zero. The indegree and out degree distributions of the APS

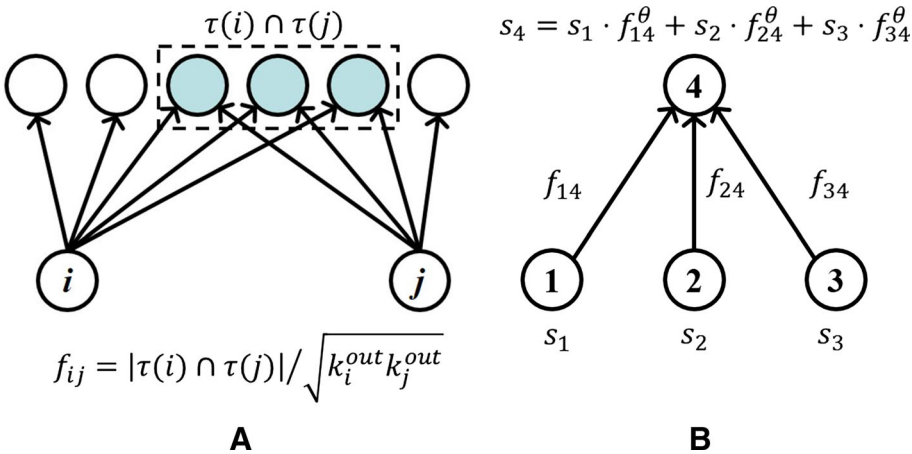

Fig. 1 (Color online) The illustration of the SPRank algorithm. **a** The calculation of the cosine index. $|\tau(i) \cap \tau(j)|$ is the number of co-cited papers (or called common neighbors) of two particular papers i and j in the citation network. It is then used as the numerator of the Cosine similarity index (see the equation of f_{ij} in this figure). **b** The aggregation process of the score from the downstream papers

citation network are shown in the Supplementary Information (SI). The ranking performance of SPRank with the incoming link is even lower than PageRank. This is why we use cosine metric based on nodes' outgoing links for the SPRank method. We also applied some other similarity indexes such as Jaccard, Hub-promoted-index, Hub-suppressed-index in our method (Lü and Zhou 2011). They all present similar results as the cosine index.

Results and discussion

Basic properties of the SPRank method

To begin our analysis, we study the basic properties of the SPRank method. As mentioned above, we will make use of the APS citation network to examine the performance of SPRank in the following analysis. One of the basic properties is the convergence of the method. As there is nonlinearity in the formula, the convergence of the method cannot be directly proved by the Perron–Frobenius theorem (Perron 1907; Frobenius 1912). Instead, we numerically investigate the convergence of our method. We first study the dependence of the total score on the iteration steps in Fig. 2a. One can see that the total score remains unchanged when $\theta = 0$. However, when $\theta > 0$ the total score first decreases with the iteration steps and then remains stable. We define a quantity *error* at step t as $\sum_i |s_i(t) - s_i(t-1)|$ where $s_i(t)$ is the score of node i at iteration step t. The results of *error* in different iteration steps are shown in Fig. 2b. If *error* can finally achieve a very small value, the ranking algorithm converges. It can be seen that SPRank converges under different θ. We also can find that our method's convergence speed increases with θ. Based on the above results, we stop the iterations when $t = 50$ in our simulation since nodes' score has reached the stable state. Another basic property of the method is the distribution of the final scores. In Fig. 2c, we investigate the distributions of the final SPRank score under different θ. We can find that the distribution of SPRank scores under different θ all roughly follows power-law form with the scaling exponents increasing with θ. Note that even though the total score decreases with the iteration steps, it can always be normalized to 1 like the classic PageRank algorithm. As we mainly focus on the ranking based on the scores, the normalization is not done in SPRank.

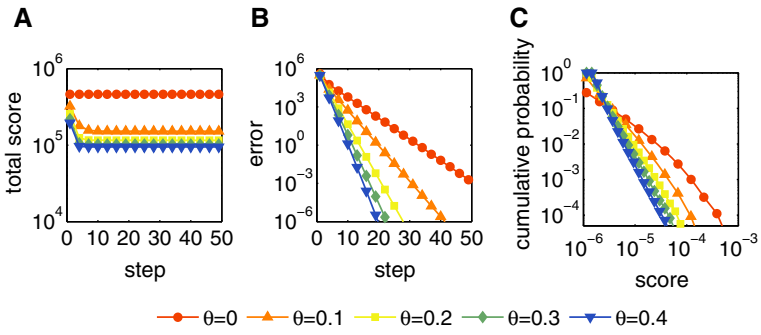

Fig. 2 (Color online) **a** The dependence of the total score on the iteration steps in SPRank, **b** the dependence of the total quantity *error* on iteration step t in SPRank and **c** the cumulative distribution of the final SPRank score. In each sub-figure we fix $c = 0.15$ but consider different θ value. Notice that SPRank reduces to PageRank when $\theta = 0$

We then study the similarity between SPRank and the other two standard methods (i.e. degree rank and PageRank). PageRank is introduced above, the Degree rank simply ranks papers according to their received citations (i.e. indegree). Specifically, we investigate the Pearson correlation coefficient and Spearman correlation coefficient between the scores of different methods. The results are shown in Fig. 3. Firstly, the score correlation between SPRank and the other two methods is shown in Fig. 3a, b, respectively. One can see the correlation between SPRank and PageRank decreases with θ. The curve starts from 1 because the SPRank reduces to PageRank when $\theta = 0$. The Pearson correlation coefficient of SPRank and PageRank stays at a low value when $\theta > 0.1$, but their Spearman correlation coefficient is higher. Normally, the highly cited papers are more important. We therefore further consider the score correlation between top-n most popular papers in Fig. 3c, d (with top-1000) and Fig. 3e, f (with top-100). We find that the Pearson correlation coefficient in these highly-cited papers are similar to the overall Pearson correlation coefficient. However, the Spearman correlation coefficient of the highly-cited papers is significantly different from the overall Spearman correlation, indicating the ranking of highly-cited papers changes remarkably by introducing the parameter θ.

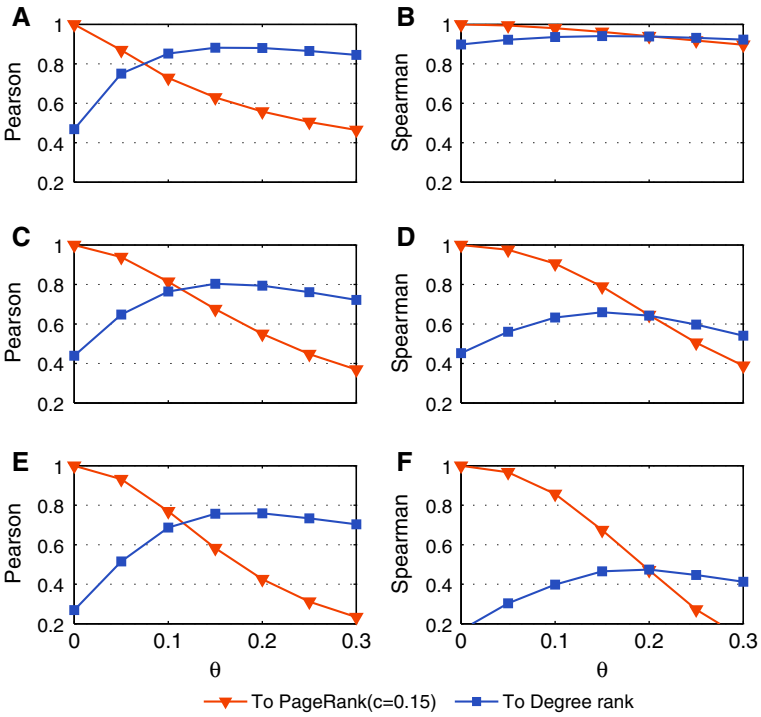

Fig. 3 (Color online) The Pearson correlation coefficient and Spearman correlation coefficient of SPRank score with indegree and PageRank score, respectively. In this figure, we fix $c = 0.15$ but different θ value to see the correlation of **a**, **b** all papers, **c**, **d** top-1000 papers and **e**, **f** top-100 papers

The effectiveness of the SPRank method

A well-performing ranking algorithm should be able to identify the truly influential nodes. As we can have access to the title of each paper, we can identify a set of these papers by selecting the Nobel prize winning papers. We regard these papers as a benchmark set to examine the performance of the PageRank and SPRank in identifying the high quality papers. The algorithm that can rank these papers higher is better. Here, we use 24 articles which author is awarded the Nobel Prize in Physics from the year 1980–2013 as our benchmark articles (see SI for details). We compare the mean rank of these papers in the PageRank and SPRank algorithms in Fig. 4. The mean rank is easy to compute. However, the result might be dominated by some lowly ranked papers. To avoid this, we propose a mean relative rank to double check the results. To obtain the mean relative rank, we first divide the rank of each Nobel prize winning paper with specific θ by the rank of this paper with $\theta = 0$, i.e. computing the ratio of the rank in SPRank and the rank in PageRank. Then this ratio of all Nobel prize winning papers is averaged to obtain the mean relative rank. If the mean relative rank decreases with θ, SPRank indeed outperforms PageRank.

In Fig. 4a, we observe that the mean rank obtains the minimum value when $\theta = 0.1$. In Fig. 4b, the mean relative rank obtains the minimum value when θ is about 0.05. θ is a tunable parameter controlling the effect of similarity in the SPRank method. When $\theta = 0$, SPRank method degenerates to the traditional PageRank method. When $\theta > 0$, an upstream node unevenly aggregates the scores from downstream nodes. The score from dissimilar downstream nodes will be suppressed. The higher the θ is, the stronger the suppressing effect is. The accuracy first increases with θ because some noisy and unreliable connections in the citation network are neglected by the SPRank method. However, as θ further increases, the score from dissimilar downstream nodes is overly suppressed, resulting in significantly loss of information in the citation network. The most extreme case is when θ is infinitely large, every upstream node only aggregates the score from the most similar downstream node. The dramatic loss of information leads to the decrease of the accuracy. Therefore, the mean rank and the mean relative rank can achieve a minimum value at a certain θ. Taken together, we could conclude that SPRank can indeed outperform PageRank in identifying these truly influential papers. However, one shouldn't set θ to a too large value. In the literature, it has been pointed out $c = 0.5$ in PageRank is a better

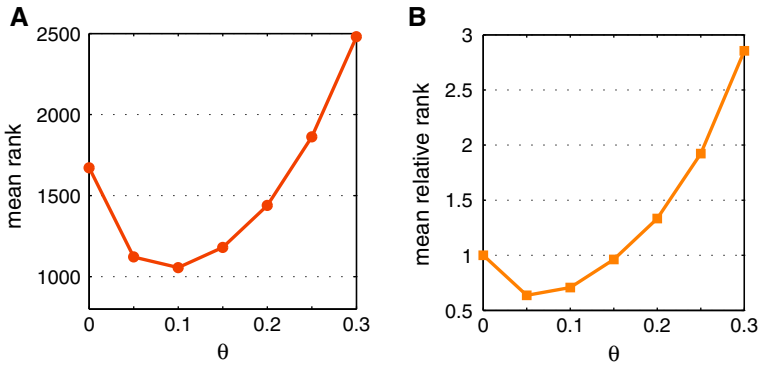

Fig. 4 (Color online) The dependence of the mean rank and the mean relative rank of 24 Nobel prize winning papers in SPRank on parameter θ in the (**a**) and (**b**), respectively

parameter for citation networks (Walker et al. 2007). Accordingly, we have also carried out the simulation in Fig. 4 with $c = 0.5$. The simulation results show that the SPRank with $c = 0.5$ performs worse than $c = 0.15$.

We further study the capability of the SPRank method in predicting the papers' future citations. We first choose a testing time t and construct the citation network based on all historical data before t. Then we calculate the PageRank and SPRank's score under different θ based on the above citation network. The aim is to see whether PageRank and SPRank's score (s) can predict the future degree growth of the papers or not. To avoid the absolute cold-start problem, we select the papers which indegree is not zero and the publication time is $[t - \Delta t_1, t]$ and then compute their future degree increment Δk in the future time $[t, t + \Delta t_2]$. We calculate the Pearson correlation coefficient between s and Δk to measure the method's prediction ability. The Pearson correlation coefficient is defined as the prediction accuracy of the SPRank method. In order to test the capability of the method in predicting small degree nodes, we also consider a set of nodes whose historical indegree is between [1, 20].

We select 20 testing time t from 1980 to 1999 with $\Delta t_1 = 3$ and different Δt_2. We use $\Delta t_2 = 3$ in Fig. 5a, b. Interestingly, like the results with the Nobel prize winning papers, there is a optimum of the Pearson correlation coefficient when varying θ. The peak exists for both node sets (i.e. nodes with nonzero historical indegree and nodes with [1, 20] historical indegree). The highest correlation coefficient is achieved when $\theta = 0.1$, indicating that the SPRank method can improve the prediction ability compared with PageRank algorithm. We then study the effect of the length of future time on predicting the future in Fig. 5c, d. The curve of SPRank in Fig. 5c, d shows its optimal results (i.e. the

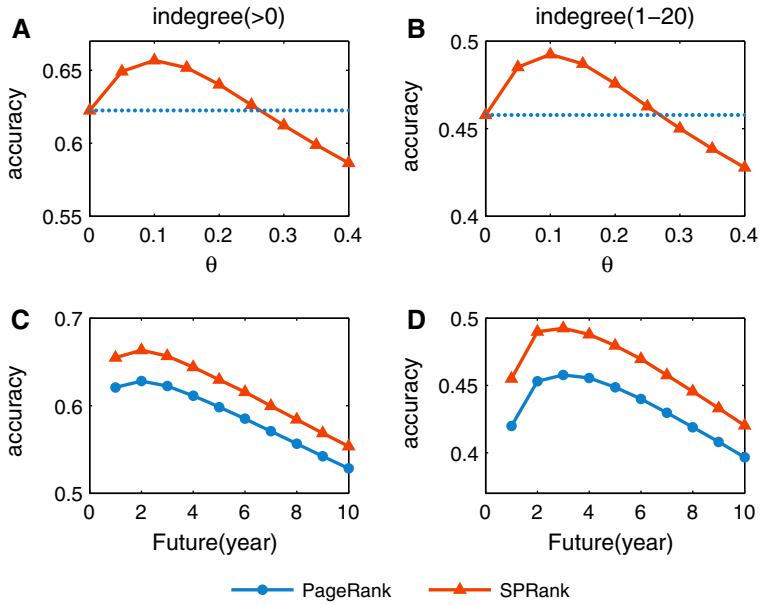

Fig. 5 (Color online) The correlation between the papers' scores at time t and their future degree increment $\langle \Delta k \rangle$ for all papers (not include papers which indegree is zero) in (**a**, **c**) and the papers of the small indegree in (**b**, **d**). We choose 3 years as future time interval in (**a**, **b**) and select different time interval in (**c**, **d**). The above results are averaged over 20 testing times from 1980 to 1999

results when $\theta = 0.1$). One can see that SPRank constantly outperforms PageRank in prediction.

The robustness of the SPRank method

In citation networks, it is common that some researchers try to manipulate the citations by self-citation or by deliberately citing his/her friends' papers. For instance, some researchers might deliberately cite their own papers when they publish new papers to push up the influence of their old publications. Consequently, self-citations are found to be very often in citation networks (Ioannidis 2015). Moreover, some irresponsible authors might carelessly select the references in their papers, resulting in assigning citations to inappropriate papers. These behaviors may exert an obvious effect on the ranking of scientific publications. Therefore, it is necessary for the designed ranking algorithm to be robust against these malicious behaviors. We further investigate the robustness of the SPRank algorithm against malicious manipulations.

Firstly, based on the APS data from 1893 to 2009, we construct a true citation network. Then we randomly select one paper with indegree $k = 0$ and regard this paper as the target paper they want to push up. We add n papers with m links each into the citation network. These new papers all cite the target paper and other $m - 1$ links each will randomly connect to other nodes. The rank of the target paper in PageRank and SPRank in the initial network are denoted as $R_p(N_0)$ and $R_s(N_0)$, respectively. At the same time, we define the rank of the target paper in PageRank and SPRank as $R_p(N_h)$ and $R_s(N_h)$ after adding hth

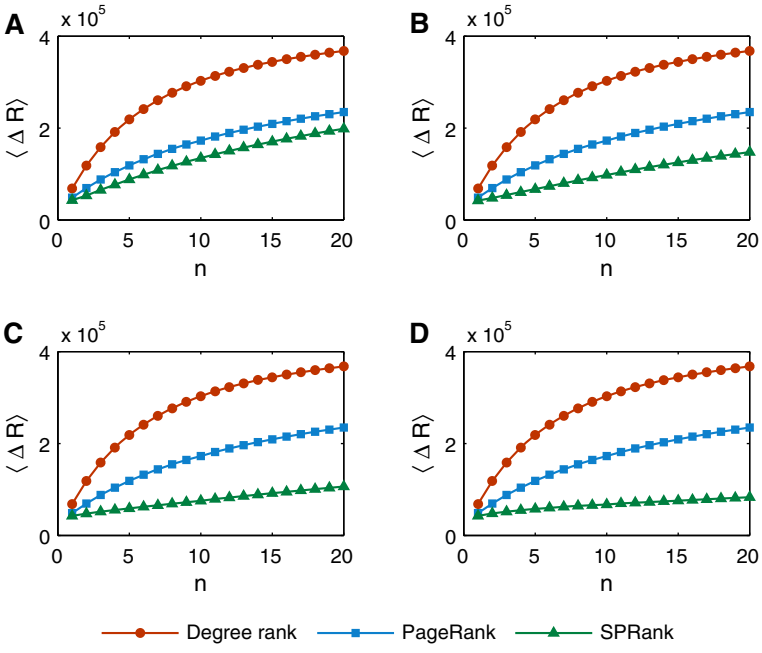

Fig. 6 The average rank change $\langle \Delta R \rangle$ of the manipulated papers in different algorithm. n is the number of new papers added and each new paper has $m = 20$ links. θ of SPRank is different in each *panel*: **a** $\theta = 0.1$, **b** $\theta = 0.2$, **c** $\theta = 0.3$ and **d** $\theta = 0.4$. Results are obtained by averaging 100 times of independent realizations

new paper into the citation network. We use ΔR_p and ΔR_s to express the rank change for PageRank and SPRank where $\Delta R_p = R_p(N_0) - R_p(N_h)$ and $\Delta R_s = R_s(N_0) - R_s(N_h)$. Naturally, smaller rank change indicates higher robustness against manipulations. We study the robustness of the Degree rank, PageRank and SPRank with the average rank change $\langle \Delta R \rangle$ in Fig. 6. The results show that the Degree rank is the most sensitive to the manipulation and the SPRank is the best in resisting such manipulation among the three methods under different θ.

Discussion and future work

Objectively ranking the quality of scientific publication is a long-standing challenge in scientometrics. Much effort has been made in this direction. Though degree is nowadays widely used to approximate the quality of papers, its result is not fair because it disregards which papers cite the paper we want to rank. The PageRank now is commonly considered as a better metric than degree, as it takes into the global information of the network when ranking nodes. However, this method is sensitive to malicious manipulation as in each iteration step each paper aggregates the score of all its downstream nodes. In this paper, we propose a new iterative ranking algorithm which highlights the contribution of the links connecting similar nodes. We find that the new method can not only significantly improve the robustness of the resultant ranking, but also have outstanding effectiveness with respect to identifying the influential papers and predicting the future citation growth.

One important property for the SPRank is that the total resource in the iteration is not constant. This feature actually helps further enhance the robustness of the method. For instance, in a network with many spurious publications aiming at pushing up the ranking of some particular papers, the SPRank will suppress the influence of these spurious publications by diminishing their resource in the network. In the literature, it has been pointed out that the main difference between spurious connections and ordinary connections in a complex network is whether the link connects two topological similar nodes (Guimerà and Sales-Pardo 2009). As a typical type of complex network, citation networks may also have some spurious publications which randomly cite other papers or deliberately cite certain papers to push up their citations. In either case, the topological similarity between the spurious papers and their cited papers will be low (usually much lower than the topological similarity between the ordinary papers and their cited papers). As SPRank aims to suppress the score passed from the dissimilar downstream nodes, it suppresses the influence of the spurious publications but not the ordinary publications. However, if the total resource is constant, the spurious publications can still mess up the general ranking of the papers because their resource will be fully aggregated by some of the upstream nodes anyway and finally propagate in the network.

There are many extensions that could be made. Though the method in this paper can more accurately predict papers future degree than PageRank, the prediction for small or zero degree nodes is still not satisfactory. To solve this problem, one may have to combine the citation network with the scientific collaboration network. In this sense, a ranking algorithm that is suitable for multiplex network is needed. A well-designed algorithm for this kind of network may effectively solve the cold-start problem in prediction. Moreover, the current algorithm disregards the time information in the citation network. We believe that adding the time dimension to the iteration ranking algorithm will improve their performance in an evolutionary way.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grant Nos. 61374175, 61174150 and 11547188, the Young Scholar Program of Beijing Normal University (2014NT38).

References

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics*, *56*(2), 235–246.

Bergstrom, C. T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, *68*(5), 314–316.

Bergstrom, C. T., & West, J. D. (2008). Assessing citations with the eigenfactor (TM) metrics. *Neurology*, *71*(23), 1850–1851.

Bollen, J., Rodriquez, 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.

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

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, *62*(2), 236–245.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, *60*(11), 2229–2243.

Fersht, A. (2009). The most influential factors: Impact factor and eigenfactor. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 6883–6884.

Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, *6*(3), 370–388.

Fiala, D., Rousselot, F., & Ježek, K. (2008). Pagerank for bibliographic networks. *Scientometrics*, *76*(1), 135–158.

Foley, J., & Della Sala, S. (2010). The impact of self-citation. *Cortex*, *46*(6), 802–810.

Frey, B. S., & Rost, K. (2010). Do rankings reflect research quality? *Journal of Applied Economics*, *13*(1), 1–38.

Frobenius, G. (1912). Über Matrizen aus nicht negativen Elementen. *Königliche Akademie der Wissenschaften.*

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379–391.

Guimerà, R., & Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, *106*(52), 22073–22078.

Ioannidis, J. P. (2015). A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. *Journal of Psychosomatic Research*, *78*(1), 7–11.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, *390*(6), 1150–1170.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing and Management*, *44*(2), 800–810.

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

Nykl, M., Ježek, K., Fiala, D., & Dostal, M. (2014). PageRank variants in the evaluation of citation networks. *Journal of Informetrics*, *8*(3), 683–692.

Perron, O. (1907). Zur theorie der matrices. *Mathematische Annalen*, *64*(2), 248–263.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

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

Sorzano, C. O. S., Vargas, J., Caffarena-Fernández, G., & Iriarte, A. (2014). Comparing scientific performance among equals. *Scientometrics*, *101*(3), 1731–1745.

Su, C., Pan, Y., Zhen, Y., Ma, Z., Yuan, J., Guo, H., et al. (2011). PrestigeRank: A new evaluation method for papers and journals. *Journal of Informetrics*, *5*(1), 1–13.

Walker, D., Xie, H., Yan, K. K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(06), P06010.

Yan, E. (2014). Topic-based PageRank: Toward a topic-level scientific evaluation. *Scientometrics*, *100*(2), 407–437.

Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information processing and management*, *47*(1), 125–134.

Yao, L., Wei, T., Zeng, A., Fan, Y., & Di, Z. (2014). Ranking scientific publications: The effect of nonlinearity. *Scientific Reports*, *4*, 6663.

Zeng, A., & Cimini, G. (2012). Removing spurious interactions in complex networks. *Physical Review E*, *85*(3), 036101.