SCIENCE CHINA
Technological Sciences

•Review•

March 2021 Vol.64 No.3: 451–461
https://doi.org/10.1007/s11431-020-1683-2

A survey on network node ranking algorithms: Representative methods, extensions, and applications

LIU JiaQi[1], LI XueRong[2*] & DONG JiChang[1]

[1] *School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China;*
[2] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Received June 8, 2020; accepted June 28, 2020; published online October 14, 2020

The ranking of network node importance is one of the most essential problems in the field of network science. Node ranking algorithms serve as an essential part in many application scenarios such as search engine, social networks, and recommendation systems. This paper presents a systematic review on three representative methods: node ranking based on centralities, PageRank algorithm, and HITS algorithm. Furthermore, we investigate the latest extensions and improvements of these representative methods, provided with several main application fields. Inspired by the survey of current literature, we attempt to propose promising directions for future research. The conclusions of this paper are enlightening and beneficial to both the academic and industrial communities.

complex networks, node ranking methods, PageRank, HITS, algorithms

Citation: Liu J Q, Li X R, Dong J C. A survey on network node ranking algorithms: Representative methods, extensions, and applications. Sci China Tech Sci, 2021, 64: 451–461, https://doi.org/10.1007/s11431-020-1683-2

1 Introduction

In the third decade of the 21st century, the world has almost reached the age of the internet of everything (IoE) [1]. In addition to the Internet, which constitutes every aspect of people's lives, many things in the real world exist in the form of complex networks, such as social networks, ecological networks, and transportation networks. Network science has also become a very important research field today. Many research findings help people to better understand the composition, characteristics, and development of the complex network world.

In the research field of network science, the ranking of network node importance is one of the most essential problems. In reality, many network-related application scenarios depend on the importance ranking of network nodes. The most famous example is the Google search engine, which ranks search results according to the importance of the page associated with the keywords. In social networks, messages from the accounts of the most influential public celebrities can be quickly propagandized to the whole network, so identifying the influential nodes is of great significance to the monitoring of public opinions and commercial public relations. In addition, the ranking of network nodes plays an important role in the fields of recommendation algorithms, scientific research evaluation, financial markets, and ecosystem protection.

In recent years, the widely used ranking algorithm of network nodes is PageRank algorithm, which constitutes the core module of Google search engine. In 2006, PageRank was selected as one of the top ten classical algorithms in data mining field by the IEEE international conference on data mining (ICDM) [2]. Another commonly used algorithm is the hyperlink-induced topic search (HITS) algorithm pro-

*Corresponding author (email: lixuerong@amss.ac.cn)

posed by Kleinberg [3], which uses the indices of two dimensions, authorities, and hubs, to measure the importance of network nodes. In addition, there are methods based on network centralities [4], node deletion [5], and mutual information [6].

Over the past twenty years, scholars have been constantly improving existing algorithms as well as proposing novel algorithms, and the academic literature related to the ranking of network nodes has continued to grow. Nowadays, some literature reviews in related fields have been published in academic journals [7]. However, since the publication date has been several years ago, the latest research findings in this field cannot be included, and the prospect of this research field is still incomplete.

This paper presents a systematic review on the representative approaches of network node ranking algorithms as well as the related latest research frontiers. Specifically, we divide the mainstream network node ranking methods into three categories, and investigate the latest improvements and applications of these representative methods. Based on the survey of current literatures, we attempt to propose some promising directions for future research. The conclusions of this paper are enlightening and beneficial to both the academic and industrial communities.

The rest of this paper is arranged as follows. Section 2 introduces the ideas and principles of three categories of representative approaches in detail. Section 3 reviews some improvement algorithms and the latest research findings based on representative methods. Section 4 presents several main application fields of network node ranking algorithms. Finally, conclusions and several promising future research directions in this field are presented in Section 5.

2 Representative methods and algorithms

In this section, we divide the mainstream network node ranking methods into three categories: node ranking methods based on centralities, the PageRank algorithm, and the HITS algorithm. In each subsection, we introduce the principles of each category, some improvements, and several applications of these representative methods.

2.1 Node ranking based on centralities

The most intuitive methods for measuring the influence of a node in the network are based on centralities. These methods measure the influence of a node by the valuation of neighbor nodes. A representative method is Degree centrality (DC) [8], which is the simplest indicator for characterizing influential nodes. DC is defined as the degree of a node i divided by the maximum number of possible connections with other nodes. The definition of DC is as follows:

$$\mathrm{DC}(i) = \frac{k_i}{n-1}, \tag{1}$$

where k_i is the degree of the node i; and n is the total number of nodes in the network.

DC ignores the global network structure and the influence of the surrounding nodes, therefore, in several cases, it is not sufficiently accurate. Other centrality measures combine the global and local influential of nodes in complex network. For example, one of the assumptions of Betweenness centrality (BC) is that the information flow propagates along the shortest path [4]. The node importance is measured by the number of shortest paths passing through it. Computation of BC requires calculating the shortest path length between each pair of nodes and recording them at the same time. Therefore, BC computation is not always efficient and scalable in real applications. Closeness centrality (CC) computes the relative distance between each pair of nodes, and rank node importance [4]. It improves the capability of BC by avoiding node deletion method and alleviating the complexity of directly calculating.

One of the limitations of centrality-based approaches is the ignorance of node position. If a node is at the core position of the network, even if the degree is small, it usually has a high influence. Furthermore, the importance of neighbour nodes should also be considered besides the centralities.

2.2 The PageRank algorithm

In 1996 at Stanford University, the PageRank algorithm was proposed by Larry Page and Sergey Brin. The algorithm was the basis on which they co-founded Google [9]. In the late 1990s, Google distinguished itself from other search methods by PageRank algorithm and had obvious advantages. It prioritizes web pages in order of importance, putting the most useful (“good stuff”) pages at the front, making it easier to get the requiring information quickly.

The basic ideas of PageRank algorithm are as following aspects: (1) The page which is linked by important pages must be an important page. When the web page A links to the web page B, it is regarded as a voting from web page A to web page B. The importance of the page is evaluated according to the number of votes (the number of links). (2) The voting of the page should be weighted by the importance of the votes (links). If a page with high PageRank value links to another page, the PageRank value of the linked page should correspondingly increase.

The principle of PageRank algorithm is summarized as the following two steps: (1) give each page a PageRank (PR) value; (2) continue to iterate through (voting) algorithm until a stable distribution is achieved. There are three ways to improve PR value: (1) the input degree; (2) the incoming links come from important pages (with high PR value); (3) the number of links in the link source page. Based on these

three points, the definition of PageRank (PR) value is as follows:

$$\mathrm{PR}_k = \sum_{j \in B_k} \frac{\mathrm{PR}_j}{L_j}, \tag{2}$$

where PR_k is the PR value of web page k; B_k represents the collection of all the pages linked to page k; L_j represents the number of outbound links (outgoing degrees) of web page j.

The equation can be transformed into a matrix formation as follows:

$$\mathbf{Apr} = \mathbf{pr}. \tag{3}$$

The problem then translates to finding the eigenvector **pr** of the matrix **A** corresponding to the eigenvalue λ=1, and rank the components of the **pr** to get the ranking of the relevant pages. Figure 1 presents an example of PageRank ranking.

The PageRank algorithm uses the output probability distribution to reflect the probability of someone randomly clicking on a web page. Before the initial calculation, the total probability will be evenly distributed to each page, so that the probability of each page being accessed is the same. Then, in the iteration, the algorithm will constantly adjust the PageRank value of each page to make it closer to the final theoretical value (i.e., convergence). Ref. [10] suggests that the convergence speed in the PageRank algorithm is not the same. Most web pages with low PR value can converge to the final PR value fast, while those with high PR value will be relatively slow. Ref. [10] also puts forward two algorithms, the Adaptive PageRank and Modified Adaptive PageRank, to significantly improve the PageRank computation speed.

Ever since the PageRank algorithm is proposed, it has been applied in various fields. Ref. [11] presents applications of PageRank in different areas. For example, PageRank is used for studying the change of the molecules of hydrogen health connection network [12]. In addition, ref. [11] also introduce the extensions of PageRank in biology, neural network science, complex systems, literature metrology, social networking, and recommender systems.

The main limitation of PageRank is that the algorithm cannot make specific recommendations according to the user's personal interests. Another limitation is that PageRank assumes that all links pointing to a page are of equal importance. However, some links are more important than others in real applications. That is, the network links should be weighted according to their importance.

2.3 The HITS algorithm

HITS algorithm is proposed by ref. [3]. HITS algorithm not only considers the number of links itself, but also the authority of links to web pages. The algorithm defines two indicators, authority and hub, to measure the importance of network nodes. If a web page has high authority value, there are many pages point to it.

Meanwhile, if a page is pointing to a large number of authoritative web pages, it has high hub value. Therefore, the authority value and the hub value are related with each other. The authority value of a web page is equal to the sum of the authority values of all pages pointing to that web page, while the hub value of a web page is the sum of the authority values of all pages that web page is pointing to. Thus, the indices of authority $a(p)$ and hub $h(p)$ can be defined as

$$a(p) = \sum_{q \in P_{\mathrm{to}}} h(q), \tag{4}$$

$$h(p) = \sum_{q \in P_{\mathrm{from}}} a(q). \tag{5}$$

HITS algorithm is calculated based on a small set of web pages. First, relevant web pages are found through text search, and the root set is obtained. Then, the web pages directly connected with these root sets are found to obtain the base set. Figure 2 presents an example of expending a root set into a base set.

HITS algorithm not only provides network node ranking, it also helps to understand authority nodes on different domains. Ref. [13] uses HITS algorithm to identify the authoritative position of traditional media in today's news media network, leading to that traditional media still play a dominant role in current digital media era. In addition, ref. [14] proposes a HIT-PR-HHBLITS prediction method for

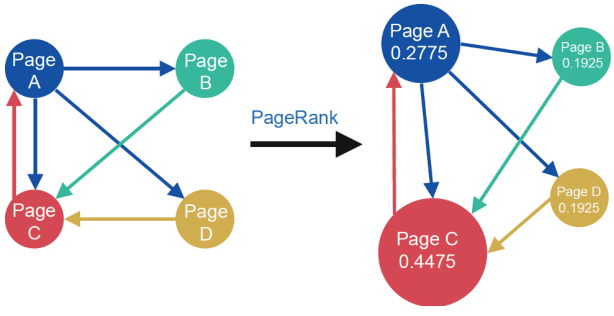

Figure 1 (Color online) An example of PageRank ranking.

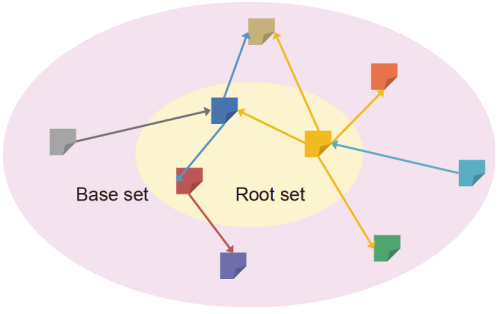

Figure 2 (Color online) An example of expending a root set into a base set.

protein remote homology detection by combining PageRank with HITS, which further improved the prediction performance. Ref. [15] uses an improved HITS algorithm to distinguish the spam and uninvited bloggers in the social network recommendation system from the real experts in the field, so as to ensure the reliability of the recommendation results.

The limitation of HITS mainly lies on the phenomenon of "topic drift". The problem appears when tightly-knit communities exist, where the authorities and hubs of these nodes are mutually reinforced. Under such circumstances, the algorithm tends to assign high importance values to nodes in tightly-knit communities and deviates from the search topic.

2.4 Linkages and differences of three representative methods

The centralities of network nodes are the most intuitive measures for influence by valuing the importance of neighbour nodes. PageRank and HITS algorithm borrows the basic idea of node centralities, and develop their own strategies for importance ranking. The PageRank algorithm values node influence based on centralities weighting by the importance of neighbour nodes. Whereas the HITS algorithm considers both authorities and degrees of neighbour nodes. In a word, node centralities provide foundation for PageRank and HITS.

HITS and PageRank algorithms are similar, since both of which consider the authority of nodes. However, the difference between the two is that (1) HITS calculates the authority value and hub value of each page and considers them separately, while PageRank only calculates the PR value; (2) HITS only deals with the collection of web pages related to keywords, and is calculated on a small set. PageRank is a global computation, which will calculate the PR value of all web pages in the Internet. Ref. [16] has made a comparative analysis of different link analysis methods, including PageRank, Weighted PageRank (WPR), HITS and CLEVER algorithm.

3 Extensions and improvements on representative methods

In this section, we present some extensions and improvements on three categories of representative methods, as well as the latest research frontiers in this field. Table 1 summarizes all the extensions of algorithms, which will be introduced in detail in the following subsections.

3.1 Improvement methods based on centralities

3.1.1 K-shell decomposition

One of the limitations of centrality-based approaches is the ignorance of node position. If a node is at the core position of the network, even if the degree is small, it usually has a high influence. Analogically, the node with large degrees but at the edge of network often has limited influence. Based on this perspective, ref. [17] puts forward *k*-shell decomposition to determine the location of the nodes in the network. It strips off the peripheral nodes on the outer layers, while nodes in the inner layer are of high influence. This method can be seen as a kind of coarse graining sorting method based on node degree. Specific decomposition process is as follows. If there are some nodes with the degree of 1 in a network, they are the least important nodes, and the connections of these nodes are removed. The rest of the network will appear some new nodes with the degree of 1, then all the edges of these nodes are removed. It repeats these steps until the rest of the network doesn't exist the nodes with the degree of one. At this point, all removal nodes form a layer, which is called 1-shell (notes for k_s=1). Continue to strip off the remaining nodes with the degree of 2 in the network, and repeat these operations until there are no nodes in the network. Figure 3 illustrates an example of *k*-shell decomposition.

In the *k*-shell decomposition, all the links are regarded equally without any weights. Ref. [18] proposes a modified

Table 1 Extension algorithms of representative methods

Representative methods	Extensions and improvement methods	Principles	Refs.
Node ranking based on centralities	*K*-shell Decomposition	A node at the core position of the network usually has a high influence	[17–19]
	Gravity centrality	The concept of universal gravitation formula of Isaac Newton to characterize the nodal influence	[20–22]
The PageRank algorithm	Personalized PageRank	Consider the user's personal interests and feedback	[23–26]
	Weighted PageRank	Assign larger rank values to more important (popular) pages	[27–30]
	LeaderRank	The probability of inputting a URL to visit other web pages is not equal	[31–33]
The HITS algorithm	HillTop	Extracts the expert pages which are related to the subject, then rank the pages linked to the expert pages	[34]
	SALSA	Divide the pages into a bipartite graph, which consists a hub set and an authority set	[35,36]

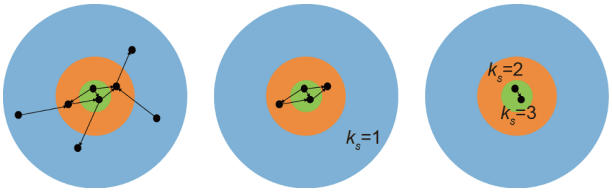

Figure 3 (Color online) An example of *k*-shell decomposition.

method taking the sum of the degrees of the end nodes as the weight of the edge. This method measures the importance of an edge by the degree of its end nodes. The modified *k*-shell decomposition algorithm is used for computing user influence on Twitter, which assigns logarithmic *k*-shell values to users, producing a measure of users that is surprisingly well distributed in a bell curve [19].

3.1.2 Gravity centrality

Based on the concepts of *k*-shell method, ref. [20] propose a new measurement of node influence called the gravitational centrality (GC). This model is similar in formation with the universal gravitation formula of Isaac Newton. The gravity is defined as the product of the *k*-shell values divided by the shortest path lengths between the two nodes. Then, GC is calculated by summing up the gravities of nodes within *n* steps (i.e., $d_{ij} \leq n$). Specifically, gravitational centrality ($G(i)$) is defined as

$$G(i) = \sum_{j \in \psi_i} \frac{k_s(i)k_s(j)}{d_{ij}^2}, \tag{6}$$

where $k_s(i)$ is the degree of node i; d_{ij} is the shortest path distance between node i and node j; ψ_i is the local nodes whose distance to node i is less than or equal to a given value r.

Later on, several improvements of GC are proposed based on extensions of gravity formula. Ref. [21] present a novel metric called Logarithm gravity (LG) centrality. In this model, the mass for each node is the degree centrality, and the distance is the length of the shortest path between a pair of nodes. Based on the study of [21,22] proposes the mixed measure of GC combining the *k*-shell value and other centrality indices (DC, BC, CC, etc.).

3.2 Improvement methods based on PageRank algorithm

3.2.1 Personalized PageRank

The original PageRank algorithm computes the PR value before the user queries, and has no relationship with the user's interests and feedback. With the development of personalized search engine, the algorithm cannot make specific recommendations according to the user's query characteristics. For this deficiency, personalized PageRank algorithm based on personal interests and feedback is proposed. On the basis of the original PageRank algorithm, user interest values and relevant feedback values are added to establish a search behavior record for users, so as to provide accurate search for specific users when they query again.

In the conceptual model, the random walk jumping behavior in the biggest difference between personalized PageRank and global PageRank. In order to reflect the user's preference, the access probability of each node in the random walk is not totally randomly selected to any node. Instead, the personalized PageRank algorithm requires the user can only jump to some specific nodes representing the user's preference. Therefore, in a stable state, users' preferred nodes and related nodes can always get a high probability of access. The random walk model can be formally expressed as

$$r = (1-\alpha)Mr + \alpha\mathbf{v}, \tag{7}$$

where α is jump probability, also known as the damping factor; M is the adjacency matrix of the network; $\mathbf{v}$ is the user preference vector (or personalization vector), and $|\mathbf{v}|=1$. If there are k nodes representing the user's preference, the sum of the values of these k nodes in $\mathbf{v}$ is one, while the values of other nodes are zero. The solution of eq. (6) is the personalized PageRank vector (PPV) corresponding to the preference vector $\mathbf{v}$, which reflects the importance of each node in the network for a given preference vector.

In practical applications, both online and offline computation of PPV require a great deal of overhead. Therefore, a lot of studies are carried out on how to improve the efficiency of PPV computing. For example, ref. [23] proposed topic-sensitive PageRank, which takes the top 16 topics in the Web open directory (ODP) as the personalized target, that is, to calculate a personalized PageRank vector r_i for each Topic T_i. Each element of v_i represents whether the corresponding page i belongs to T_i. For pages that do not belong to T_i, the value is 0, while pages belonging to T_i have equal values.

Topic-sensitive PageRank is one of the coarse-grained computation methods of the PPV, in which any size of Web data is mapped into a limited several topics or dimensions, thus improving the offline procession and computational efficiency. The PPR algorithms using similar ideas include the BlockRank method proposed by ref. [24]. However, this method cannot guarantee high accuracy. Some researchers put forward the method of reducing the accuracy of PPV calculation to realize the scalable personalized PPV, that is, the approximation method of PPV. For example, ref. [25] propose the Web Skeleton method, which defines that hub nodes are nodes with a large number of links in the Web, so most possible paths between two nodes need to pass through one or more hub nodes. Based on a reasonably selected set of hubs, the sum of the probabilities of hub paths is close to the

sum of the probabilities of all paths. Therefore, for any preferred node q, the partial probability that goes through the hub node from q can be used as the estimation of the real probability. Another way is a method proposed by ref. [26], which is called fingerprint. It uses random walk step model to sampling for all paths, and the sum of the probability of PPV of the sampled paths (sampled tours) is estimated as the real value. An overview article about personalized PageRank [37] summarizes five PPR calculation methods in recent years: direct equation solving [38], iterative equation solving [39], bookmark coloring [40], dynamic programming [41] and Monte-Carlo sampling [42].

Recent years, the PPR algorithm combined with top-k query technology [43] is the latest research frontier in this field, which is also known as single-source PPR (SSPPR). Ref. [44] proposes kPAR algorithm, which implements real-time Top-k PPR computation. The kPAR designs novel algorithms combining adaptive forward push and inverted random walks, while utilizes load balancing to realize the capability of GPUs. Ref. [45] combines two existing methods—Forward Push and Monte Carlo Random Walk to realize an approximate SSPPR solution, as well as a module for top-k selection with high pruning power.

3.2.2 *Weighted PageRank algorithm*

Standard PageRank assumes that all links pointing to a page are of equal importance. However, some links are more important than others in real applications. One of the extensions of PageRank algorithm is Weighted PageRank Algorithm. Instead of dividing the importance value evenly among its out-link nodes, this algorithm assumes that more popular nodes have larger rank values [27]. That is, the node importance is proportional to its number of in-links and out-links. The number of in-links and out-links of a node is recorded as $W^{\text{in}}_{(v,u)}$ and $W^{\text{out}}_{(v,u)}$, respectively, which are defined as follows:

$$W^{\text{in}}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p}, \tag{8}$$

$$W^{\text{out}}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p}, \tag{9}$$

where I_u and I_p represent the number of in-links of page u and page p, respectively. O_u and O_p represent the number of out-links of page u and page p, respectively. $R(v)$ denotes the reference page list of page v.

Ref. [28] incorporates user browsing behavior in Weighted PageRank algorithm based on number of visits of links (VOL), and develops a new algorithm of Weighted PageRank. In this algorithm, the node importance is proportional to the VOL of out-links. However, it only considers the number of in-links for population, and fails to identify the popularity of outgoing links, which was incorporated in Weighted PageRank algorithm by ref. [29]. Furthermore, ref. [30] proposes a modified ranking mechanism considering both the contents as well as visits of links.

3.2.3 *LeaderRank algorithm*

In PageRank algorithm, the random jump probability of each node is the same, that is, starting from any web page, the probability of inputting a URL (uniform resource locator) to visit other web pages is equal. But in reality, when people are browsing the popular web pages with rich contents (the nodes with large degree), the probability of users choosing to use the address bar is far less than the boring website (the nodes with small degree). Instead of using jumping probability of PageRank algorithm, LeaderRank algorithm adds a background node and constructs the two-way edges between the background node and all other nodes [46]. In LeaderRank algorithm, the probability of entering a URL from a page to visit next page is the equivalent of the probability of the page visiting background nodes. The probability is negatively related to the number of links on the page. The more the link from the local access, the less the probability. One unit of initial LeaderRank value is given to all nodes other than the background node v_g, namely $\text{LR}_i(0)=1, \forall\ i \neq g$; $\text{LR}_g(0)=0$, followed by the following iterative process until steady state:

$$\text{LR}_i(t) = \sum_{j=1}^{n+1} \frac{a_{ji}}{k_j^{\text{out}}} \text{LR}_j(t-1). \tag{10}$$

In steady state, the LeaderRank value of the background node, $\text{LR}_g(t_c)$, is divided equally to the other n nodes, so the final LeaderRank value of node v_i is obtained:

$$\text{LR}_i = \text{LR}_i(t_c) + \frac{\text{LR}_g(t_c)}{n}. \tag{11}$$

In the standard LeaderRank algorithm, the connection between the background node and all nodes is the same. Many scholars have proposed an improved weighted LeaderRank, which believes that when visiting other nodes from the background node, the nodes with a large degree should have a higher probability to be visited [31–33].

3.3 Improvement methods based on HITS algorithm

3.3.1 *HillTop algorithm*

HillTop algorithm is a link analysis algorithm proposed by Krishna Baharat and George A. Mihaila of the University of Toronto in 2000, and is later given to Google for algorithm upgrading. HillTop integrates the ideas of HITS and PageRank [34]. Similar to HITS, it is a link analysis algorithm for user query requests. Based on the user query, a high-quality collection of web pages can be obtained. Meanwhile, in the process of weight propagation, HillTop determines the ranking weight of search results by the number and quality of page entries.

Figure 4 presents the framework of HillTop algorithm.

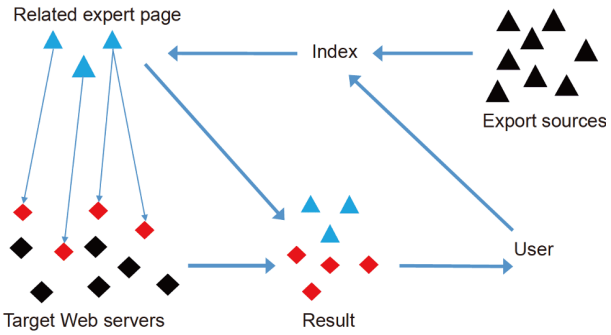

Figure 4 (Color online) Framework of HillTop algorithm.

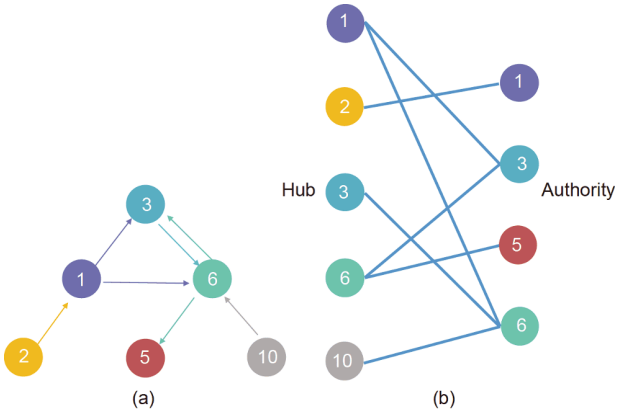

Figure 5 (Color online) The origin graph (a), and the bipartite graph (b).

HillTop algorithm firstly extracts the expert pages from a vast amount of web pages as a subset, and establish the index. The so-called expert page is an essential definition in HillTop, namely high-quality pages related to the subject. The links from expert pages are non-affiliated pages with each other (two pages do not belong to the same site). When the HillTop receives the user's query request, it selects some pages from expert pages which is strongly correlated with the user's query subject. Then the target pages are ranked according to the link relationship between the target pages and the relevant expert pages, and finally the relevant expert pages and the target pages with high scores are output as the result.

3.3.2 SALSA algorithm

Stochastic approach for link structure analysis (SALSA) was proposed by ref. [35]. Similar to HillTop, this algorithm combines the ideas of HITS and PageRank. It is similar to the HITS algorithm, which assigns a hub and authority value to a web page, and is also created through the Markov chain like PageRank. Different from global calculation of PageRank but similar to HITS, it selects highly relevant web pages after receiving the user's query request, as a "root set", and will assign the pages which are directly linked with the root set as "expand set". Finally, it determines the rankings based on link analysis within the expand page set.

After getting the page ranking, the algorithm will divide the pages in expand set into a bipartite graph, which consists a hub set and an authority set. The partitioning rules can be summarized as follows: if a web page contains out-degree, then the web page is placed into the hub set; if a web page contains in-degree from another web page, the web page will be placed in the authority set; if a web page includes both out-degree and in-degree, it can be placed in both sets. Finally, the output from the hub set constitutes the edge of the bipartite network. For example, Figure 5(a) and (b) present an example of dividing the set of Figure 5(a) into a bipartite network. However, different from HITS, the original directed links are transformed into undirected links after converting to a bipartite graph, and other partition steps are exactly the same as HITS. Therefore, SALSA algorithm can guarantee the relevance of user queries, thus avoiding the problem of topic drift. After subsets are subdivided, SALSA does not adopt authority-hub mutual enhancement method of HITS, but adopts PageRank's random walk.

Combining PageRank algorithm, SALSA solve the problem of search topic drift of HITS. Furthermore, ref. [35] also reveals a TKC (tightly knit community) effect of HITS. TKC is a small but highly connected community. During the computation of HITS, pages belonging to TKC community are tending to obtain underestimated authority values, or topic drifts. SALSA algorithm solves the TKC effect. Ref. [36] compare PageRank, HITS and SALSA, and discover the poor performance of HITS and SALSA in some networks, then put forward the improvement measures.

4 Applications of node ranking approaches

4.1 Social networks

Node ranking algorithms are applicated in social networks to calculate the importance of nodes and identify opinion leaders, or "influential spreader" [47–49]. Ref. [50] proposes a novel framework, motif-based PageRank (MPR), to incorporate higher-order relations into the conventional PageRank computation for user ranking in social networks. Based on social meta path, ref. [51] proposed an algorithm to identify the most influential individuals in complex online social networks. PageRank algorithm is also used for estimating approximations in calculations of non-polynomial complexity. In studies such as ref. [52], the proposed Influence Rank measure is estimated using PageRank.

Another application in the field of social networks is discovering opinion formation process. Ref. [53] proposes the rank model of opinion formation and investigate its rich properties on real directed networks of several universities. A related research problem is influence maximization in social networks. Ref. [54] applies the PageRank in signed

social networks to study influence maximization in online social networks (OSNs), where the relations include hostile ones except friendly ones. Ref. [55] develops a quantitative metric, named Group-PageRank, to quickly estimate the upper bound of the social influence for social influence maximization. More generally, ref. [56] provide a focused study on understanding of PageRank as well as the relationship between PageRank and social influence analysis. They show that the authority computation by PageRank can be enhanced with more generalized priorities.

4.2 Recommender systems

The enrich online data nowadays provide an opportunity of the popularity of data-intensive applications such as recommender systems. In particular, the notion of similarity between items is one of the core modules of content-based recommender systems. Therefore, the efficiency of a recommendation engine is heavily relied on the right similarity metric of complex networks [57]. The node ranking algorithm is then applied for recommender systems to improve similarity measurement between entities. For example, ref. [58] incorporates userrank as weight of a user based on PageRank into item similarities computing. Ref. [59] proposes a Topical PageRank based algorithm for recommender systems, which aim to rank products by the analysis of previous user-item relationships, and recommend top-rank items to potentially interested users.

In addition to measure similarity, the node ranking algorithm is also integrated into some novel methods to capture graph structures and improve the performance of recommender systems. Ref. [60] presents a novel personalization algorithm which combines usage data and link analysis techniques that taking also into account the Web structure for ranking and recommending. In the recent paper of ref. [61], the proposed algorithm finds graph clusters using the personalized PageRank vectors to determine a set of clusters, which are used for building a recommendation system that can recommend products based on the buying behaviour of the users. Based on above perspectives, node ranking algorithms are also applied in some industrial-specific recommender systems such as mobile apps recommendation [62], academic resources recommendation [63], manufacturing service recommendation [64] and expert recommendation [65].

4.3 Citation networks

In the fields of scientometrics and bibliometrics, the PageRank algorithm provides an alternative method for measuring the importance of scientific papers or authors besides traditionally used citation analysis [66]. As ref. [67] states, a meaningful advantage of PageRank is that it could largely eliminate the flattery of academic influence caused by author self-citations. Therefore, node ranking algorithms are applied to evaluate influence of articles and authors by ranking node importance of citation networks. Multiple extensions of PageRank have been proposed, which are more suitable for author ranking. In the study of ref. [68], PageRank and its modifications are used for the evaluation of various types of citation networks. They find that the best ranking of authors is obtained by using a publication citation network from which self-citations are eliminated, which confirms the statement of ref. [67]. Other studies consider co-authorship network when discovering author impact using PageRank and HITS algorithm [69,70].

In addition to article ranking and author ranking, PageRank and HITS is also introduced to journal impact evaluation and award prediction. Ref. [71] presents a model named the multiple-link, mutually reinforced journal-ranking (MLMRJR) model based on the PageRank and the HITS algorithms that considers not only the quantity and quality of citations in intra-networks, but also the mutual reinforcement in inter-networks. Ref. [72] proposes an alternative approach, Eigenfactor Metrics, to access scholarly journals based on the PageRank algorithm. By this approach, citations from top journals are weighted more heavily than citations from lower-tier publications. Ref. [73] proposes a temporal probabilistic ranking model to predict ACM's SIG (Special Interest Groups) award which combines content with citation network analysis.

4.4 Traffic and transportation networks

Traffic analysis based on transportation networks is an important work to transport people and goods to their destinations in a quick and efficient manner. Recent years, some studies have integrated node ranking algorithms to find important spots which have a strong effect on transportation efficiency. Ref. [74] identifies the points where the traffic is congested and the intersections where the traffic goes frequently in and out between cities. [75] reports a web-based viewer of taxi probe data, and propose a traffic simulation model based on the concept of PageRank. Ref. [76] proposes two geographically modified PageRank algorithms that incorporate geographic considerations into PageRank algorithms to identify the spatial concentration of human movement in a geospatial network. Based on the traffic spots evaluation, node ranking algorithms make it possible to achieve traffic optimization of decision making [77].

In addition to discovering important traffic spots, the PageRank and other node ranking algorithm is further applied to the prediction of transportation networks. Ref. [78] predicts the situations of urban traffic network congestion with an improved PageRank model, and obtain a good forecast effect on the urban traffic congestion. Ref. [79] finds that the

PageRank values can act as signatures in predicting upcoming traffic congestions, and experimentally confirm it based on the trajectory data of 12000 taxies in Beijing city for one month.

4.5 Financial and economic networks

Recently, financial complex networks have achieved more attention from scholars of various field. Standard methods of network analysis including PageRank and HITS provide new perspectives to understanding the underlying mechanisms and driving forces in a financial market. For example, ref. [80] proposes a method based on cointegration instead of correlation to construct financial complex network in Chinese stock market, and analyze these directed, weighted and non-symmetric networks by using node importance measures. Instead of focusing on China, ref. [81] constructs cointegration networks among 26 global stock market indices, and investigate the influence rank of network nodes. PageRank is also used in ref. [82] to produce a ranking of participate banks in the Canadian Large Value Transfer System in terms of their daily liquidity holdings, explaining liquidity flows through a financial network and why observed distributions of liquidity differ from the initial distributions.

Besides applications in financial markets, node ranking algorithms and their modifications have been applied to other fields such as fraud detection and world economic systems [83]. For example, ref. [84] makes use of secure multiparty computation (MPC) techniques. In this model, each party of the transaction network only learns the PageRank values of its own accounts, and multiple parties compute the PageRank values jointly of their combined transaction networks. Ref. [85] uses the HITS algorithm to calculate hub and authority values in the world trade network (WTN) of various countries from 1992 to 2012. It also found the pattern of world trade authority changed during this period.

5 Conclusions and promising directions of future research

In this study, we have conducted a systematic survey on literatures of the node ranking approaches, and comprehensively describe its mechanism, extensions and applications. Specifically, we introduce some representative algorithms such as PageRank and HITS. The extensions of node ranking algorithms mostly focus on the problems of personalization, computational efficiency and weighted links. So far, node ranking approaches have been applied in social networks, recommender systems, citation networks, transportation networks and financial or economic networks.

According to our systematic review on the node ranking algorithms, we have proposed the following plausible research directions in the future.

(1) Improving the capacity of node ranking algorithms in big data environment. Nowadays, online applications such as web search, social networks and graph analytics are mostly under the big data environment. In addition, more and more related applications require real-time query, i.e., within less than 100 ms, on an Internet-scale graph with billions of edges. Due to the immense computational cost of node ranking algorithms and the difficulty of designing a parallelized node ranking algorithm, challenges remain to implement node ranking in big data applications [44]. How to guarantee on both result quality and worst-case efficiency of node ranking in real-time big data environment is a promising research question in the future.

(2) Integrating node ranking algorithms with spatial and temporal analysis. Current studies are not well designed to handle spatial or temporal sensitive networks. Spatial analysis is one of the important subjects in transportation networks, community detection, location-based services (LBS) and other Web applications. Meanwhile, dynamic graphs and time-evolving networks become more popular among online services, which makes the importance of nodes may change during the lifetime of the network. Extending PageRank and other algorithms which are adaptive to networks with spatial or temporal dimension will gain more attention in the future.

(3) The application of node ranking algorithms in novel forms of complex networks. According to section 4 in this paper, node ranking approaches have been applied to various types of networks and application scenarios, such as social networks, citation networks, transportation networks and financial networks. Recent years, many literatures have studied novel networks such as multiplex networks, semantic networks, ecological networks and cash flow networks. The integration of node ranking algorithms into these new forms of networks will provide new perspectives for future research.

This work was supported by the National Natural Science Foundation of China (Grant No. 71901205).

1 Kang B, Kim D, Choo H. Internet of everything: A large-scale autonomic IoT gateway. IEEE Trans Multi-Scale Comp Syst, 2017, 3: 206–214
2 Wu X, Kumar V. The Top Ten Algorithms in Data Mining. Boca Raton (FL): CRC Press, 2009
3 Kleinberg J M. Authoritative sources in a hyperlinked environment. J ACM, 1999, 46: 604–632
4 Freeman L C. Centrality in social networks conceptual clarification. Social Networks, 1978, 1: 215–239
5 Restrepo J G, Ott E, Hunt B R. Characterizing the dynamical importance of network nodes and links. Phys Rev Lett, 2006, 97: 094102
6 Tan F, Xia Y, Zhu B. Link prediction in complex networks: A mutual information perspective. PLoS ONE, 2014, 9: e107056

7 Chung F. A brief survey of PageRank algorithms. IEEE Trans Netw Sci Eng, 2014, 1: 38–42
8 Bonacich P. Factoring and weighting approaches to status scores and clique identification. J Math Sociol, 1972, 2: 113–120
9 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Comput Networks ISDN Syst, 1998, 30: 107–117
10 Kamvar S, Haveliwala T, Golub G. Adaptive methods for the computation of PageRank. Linear Algebra Appl, 2004, 386: 51–65
11 Gleich D F. PageRank beyond the Web. SIAM Rev, 2015, 57: 321–363
12 Mooney B L, Corrales L R, Clark A E. MoleculaRnetworks: An integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation. J Comput Chem, 2012, 33: 853–860
13 Majó-Vázquez S, Cardenal A S, Segarra O, et al. Media roles in the online news domain: Authorities and emergent audience brokers. Media Commun, 2020, 8: 98–111
14 Liu B, Jiang S, Zou Q. HITS-PR-HHblits: Protein remote homology detection by combining PageRank and hyperlink-induced topic search. Briefings BioInf, 2020, 1: 298–308
15 Khan U U S, Ali M, Abbas A, et al. Segregating spammers and unsolicited bloggers from genuine experts on Twitter. IEEE Trans Dependable Secure Comput, 2016, 4: 551–560
16 Jain A, Sharma R, Dixit G, et al. Page ranking algorithms in web mining, limitations of existing methods and a new method for indexing web pages. In: 2013 International Conference on Communication Systems and Network Technologies. Gwalior, 2013. 640–645
17 Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. Nat Phys, 2010, 6: 888–893
18 Manuel S, Kumar K R. An improved *k*-shell decomposition for complex networks based on potential edge weights. Int J Appl Math Sci, 2016, 9: 163–168
19 Brown P E, Feng J. Measuring user influence on twitter using modified *k*-shell decomposition. In: Fifth International AAAI Conference on Weblogs and Social Media. Menlo Park, 2017
20 Ma L L, Ma C, Zhang H F, et al. Identifying influential spreaders in complex networks based on gravity formula. Physica A-Statistical Mech its Appl, 2016, 451: 205–212
21 Niu J, Yang H, Wang L. Logarithmic gravity centrality for identifying influential spreaders in dynamic large-scale social networks. In: 2017 IEEE International Conference on Communications (ICC). Paris, 2017. 1–6
22 Wang J, Li C, Xia C. Improved centrality indicators to characterize the nodal spreading capability in complex networks. Appl Math Comput, 2018, 334: 388–400
23 Haveliwala T H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Trans Knowl Data Eng, 2003, 15: 784–796
24 Kamvar S D, Haveliwala T H, Manning C D, et al. Exploiting the block structure of the web for computing PageRank. Stanford University Technical Report, 2003
25 Jeh G, Widom J. Scaling personalized web search. In: Proceedings of the 12th International Conference on World Wide Web. New York, 2003. 271–279
26 Fogaras D, Rácz B, Csalogány K, et al. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. Internet Math, 2005, 2: 333–358
27 Xing W, Ghorbani A. Weighted PageRank algorithm. In: Proceedings. Second Annual Conference on Communication Networks and Services Research. Fredericton, 2004
28 Tyagi N, Sharma S. Weighted page rank algorithm based on number of visits of links of web page. Int J Soft Comput Eng, 2012, 3: 2231–2307
29 Tuteja S. Enhancement in weighted pagerank algorithm using VOL. J Computer Eng, 2013, 5: 135–141
30 Prajapati R, Kumar S. Enhanced weighted PageRank algorithm based on contents and link visits. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). New Delhi, 2016
31 Zhang Z H, Jiang G P, Song Y R, et al. An improved weighted LeaderRank algorithm for identifying influential spreaders in complex networks. In: 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). Guangzhou, 2017. 748–751
32 Luo L, Yang Y, Chen Z, et al. Identifying opinion leaders with improved weighted LeaderRank in online learning communities. IJPE, 2018, 2: 193–201
33 Jiang S, Zhang X, Cao Z. An improved LeaderRank algorithm for identifying critical components in service-oriented systems. J Phys Conf Ser, 2019, 1213: 032012
34 Bharat K, Mihaila G A. Hilltop: A search engine based on expert documents. In: Proceedings of the 9th International WWW Conference, 2000
35 Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Comput Networks, 2000, 33: 387–401
36 Farahat A, LoFaro T, Miller J C, et al. Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization. SIAM J Sci Comput, 2006, 27: 1181–1201
37 Park S, Lee W, Choe B, et al. A survey on personalized PageRank computation algorithms. IEEE Access, 2019, 7: 163049
38 Tong H, Faloutsos C, Pan J Y. Fast random walk with restart and its applications. In: Sixth International Conference on Data Mining (ICDM'06). Hong Kong, 2016
39 Kamvar S D, Haveliwala T H, Manning C D, et al. Extrapolation methods for accelerating PageRank computations. In: Proceedings of the 12th International Conference on World Wide Web. New York, 2003. 261–270
40 Berkhin P. Bookmark-coloring algorithm for personalized PageRank computing. Internet Math, 2006, 3: 41–62
41 Sarlós T, Benczúr A A, Csalogány K, et al. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In: Proceedings of the 15th International Conference on World Wide Web. Edinburgh Scotland, 2006
42 Avrachenkov K, Litvak N, Nemirovsky D, et al. Monte Carlo methods in PageRank computation: When one iteration is sufficient. SIAM J Numer Anal, 2007, 45: 890–904
43 Soliman M A, Ilyas I F, Chang K C C. Top-k query processing in uncertain databases. In: 2007 IEEE 23rd International Conference on Data Engineering. Istanbul, 2007
44 Shi J, Yang R, Jin T, et al. Realtime top-k personalized PageRank over large graphs on GPUs. Proc VLDB Endow, 2019, 13: 15–28
45 Wang S, Yang R, Wang R, et al. Efficient algorithms for approximate single-source personalized PageRank queries. ACM Trans Database Syst, 2019, 44: 1–37
46 Lü L, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case. PLoS ONE, 2011, 6: e21202
47 Hu W, Zou H, Gong Z. Temporal PageRank on social networks. In: International Conference on Web Information Systems Engineering. Miami, 2015
48 Li Q, Zhou T, Lü L, et al. Identifying influential spreaders by weighted LeaderRank. Physica A-Statistical Mech its Appl, 2014, 404: 47–55
49 Zhang T, Liang X. A novel method of identifying influential nodes in complex networks based on random walks. J Inf Comput Sci, 2014, 11: 6735–6740
50 Zhao H, Xu X, Song Y, et al. Ranking users in social networks with Motif-based PageRank. IEEE Trans Knowl Data Eng, 2019, 1: 1
51 Le V V, Nguyen H T, Snasel V, et al. Identify influential spreaders in online social networks based on social meta path and PageRank. In: International Conference on Computational Social Networks. Ho Chi Minh City, 2016. 51–61
52 Hajian B, White T. Modelling influence in a social network: Metrics

and evaluation. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. Boston, 2011
53 Kandiah V, Shepelyansky D L. PageRank model of opinion formation on social networks. Physica A-Statistical Mech its Appl, 2012, 391: 5779–5793
54 Chen S, He K. Influence maximization on signed social networks with integrated PageRank. In: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (Smart City). Chengdu, 2015
55 Liu Q, Xiang B, Chen E, et al. Influence maximization over large-scale social networks: A bounded linear approach. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, 2014
56 Xiang B, Liu Q, Chen E, et al. Pagerank with priors: An influence propagation perspective. In: Twenty-Third International Joint Conference on Artificial Intelligence. Beijing, 2013
57 Nguyen P, Tomeo P, Di Noia T, et al. An evaluation of SimRank and Personalized PageRank to build a recommender system for the Web of Data. In: Proceedings of the 24th International Conference on World Wide Web. Florence, 2015. 1477–1482
58 Jiang F, Wang Z. PageRank-based collaborative filtering recommendation. In: International Conference on Information Computing and Applications. Tangshan, 2010
59 Zhang L, Zhang K, Li C. A topical PageRank based algorithm for recommender systems. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008
60 Eirinaki M, Vazirgiannis M. Usage-based PageRank for web personalization. In: Fifth IEEE International Conference on Data Mining (ICDM'05). Houston, 2005
61 Al_Janabi S, Kadiam N. Recommendation system of big data based on PageRank clustering algorithm. In: International Conference on Big Data and Networks Technologies, 2019
62 Zhong X, Zhang Y, Yan D, et al. Recommendations for mobile apps based on the HITS algorithm combined with association rules. IEEE Access, 2019, 7: 105572
63 Lu M, Wei X, Gao J, et al. AHITS-UPT: A high quality academic resources recommendation method. In: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (Smart City). Chengdu, 2015
64 Zhang W Y, Zhang S, Guo S S. A PageRank-based reputation model for personalised manufacturing service recommendation. Enterprise Inf Syst, 2017, 11: 672–693
65 Rafiei M, Kardan A A. A novel method for expert finding in online communities based on concept map and PageRank. Hum Cent Comput Inf Sci, 2015, 5: 10
66 Chen P, Xie H, Maslov S, et al. Finding scientific gems with Google's PageRank algorithm. J Informetrics, 2007, 1: 8–15
67 Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Inf Process Manage, 2008, 44: 800–810
68 Nykl M, Ježek K, Fiala D, et al. PageRank variants in the evaluation of citation networks. J Informetrics, 2014, 8: 683–692
69 Yan E, Ding Y. Discovering author impact: A PageRank perspective. Inf Process Manage, 2011, 47: 125–134
70 Silva J, Aparício D, Silva F. Feature-enriched author ranking in incomplete networks. Appl Netw Sci, 2019, 4: 74
71 Yu D, Wang W, Zhang S, et al. A multiple-link, mutually reinforced journal-ranking model to measure the prestige of journals. Scientometrics, 2017, 111: 521–542
72 West J D, Bergstrom T C, Bergstrom C T. The Eigenfactor MetricsTM: A network approach to assessing scholarly journals. CRL, 2010, 71: 236–244
73 Yang Z, Yin D, Davison B D. Award prediction with temporal citation network analysis. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, 2011
74 Kim Y Y, Kim H A, Shin C H, et al. Analysis on the transportation point in Cheongju City using Pagerank algorithm. In: Proceedings of the 2015 International Conference on Big Data Applications and Services. Jeju Island Republic of Korea, 2015
75 Mukai N. PageRank-based traffic simulation using taxi probe data. Procedia Comput Sci, 2013, 22: 1156–1163
76 Chin W C B, Wen T H. Geographically modified PageRank algorithms: Identifying the spatial concentration of human movement in a geospatial network. PLoS ONE, 2015, 10: e0139509
77 Pop F, Dobre C. An efficient PageRank approach for urban traffic optimization. Math Problems Eng, 2012, 2012: 1–9
78 Zhang T, Li G, Xu Y, et al. Prediction of transportation network based on PageRank algorithm. In: 2016 5th International Conference on Advanced Materials and Computer Science. Qingdao, 2016
79 Wang M, Yang S, Sun Y, et al. Discovering urban mobility patterns with PageRank based traffic modeling and prediction. Physica A-Statistical Mech its Appl, 2017, 485: 23–34
80 Tu C. Cointegration-based financial networks study in Chinese stock market. Physica A-Statistical Mech its Appl, 2014, 402: 245–254
81 Yang C, Chen Y, Niu L, et al. Cointegration analysis and influence rank—A network approach to global stock markets. Physica A-Statistical Mech its Appl, 2014, 400: 168–185
82 Bech M L, Chapman J T E, Garratt R J. Which bank is the "central" bank? J Monetary Econ, 2010, 57: 352–363
83 Cheng X, Shaoyi L S, Hua Z. Measuring the systemic importance of interconnected industries in the world economic system. Industr Mngmnt Data Syst, 2017, 117: 110–130
84 Sangers A, van Heesch M, Attema T, et al. Secure multiparty PageRank algorithm for collaborative fraud detection. In: International Conference on Financial Cryptography and Data Security. Kota Kinabalu, Sabah, 2019
85 Deguchi T, Takahashi K, Takayasu H, et al. Hubs and authorities in the world trade network using a weighted HITS algorithm. PLoS ONE, 2014, 9: e100338