Network node ranking methods: A case study with Chinese University Weblink

HU, Mingyun
Department of Mathematics
mhuae@connect.ust.hk

Abstract

Node ranking is a crucial problem in complex network analysis, as it aims to identify the most important or influential nodes in a network based on certain criteria. Centrality-based ranking, PageRank-based ranking, and HITS-based ranking are the most widely used node ranking methods. However, finding the appropriate ranking method for a given network and task can be challenging, as different methods may capture different aspects of node importance and may perform differently on different networks. In this study, we investigate the performance of various node ranking methods on the Chinese University Weblink dataset. Our results show that different methods perform differently on this dataset, highlighting the importance of carefully considering the underlying assumptions and limitations of different ranking methods when interpreting the results of network analysis.

1 Introduction

In the era of the internet of everything, complex networks have become ubiquitous in our daily lives, from social networks to transportation networks (4). The study of complex networks has become a crucial research topic, and node ranking is a key problem in complex network analysis. Node ranking aims to identify the most important or influential nodes in a network based on certain criteria. Various algorithms have been proposed to determine the importance or relevance of nodes in a network.

The most widely used node ranking methods can be classified into three categories (7): centrality-based ranking, PageRank-based ranking, and hyperlink-induced topic search (HITS)-based ranking. Centrality-based ranking methods, such as degree centrality (DC) (1), closeness centrality (CC) (3), and betweenness centrality (BC) (2), measure the importance of nodes based on their position in the network. PageRank-based ranking method assigns an initial PageRank value to each node and then uses a voting algorithm to determine the importance of each node (2). The value of PageRank represents the importance of each node. HITS-based ranking, on the other hand, ranks nodes based on both the number and authority of the links to them (5). In recent years, some extension and improvements on these three categories of representative methods, such as Weighted PageRank (WPR) (8), Stochastic approach for link structure analysis (SALSA) (6).

However, finding the proper node ranking method for a given network and a specific task is not a trivial problem, as different methods may capture different aspects of node importance and may perform differently on different networks. In this study, we investigate the performance of various node ranking methods on Chinese University Weblink dataset. The remainder of the report is organized as follows. Section 2 describes the dataset and evaluation metrics. Section 3 offers a detailed introduction to the node ranking methods. Section 4 provides a comprehensive analysis and discussion of the performance of each method. Section 5 summarizes the finding of this study.

2 Data and evaluation metrics

The dataset contains 76 universities of Chinese mainland. $W \in \Re^{76\times 76}$ is the link matrix whose (i, j)-th element gives the number of links from university i to j. ResearchRank denotes the research ranking of these universities. The universities with top-5 ResearchRank are shown in Table 1.

Table 1: Universities with top-5 ResearchRank.

ResearchRank	1	1	3	4	5
University	pku	tsinghua	fudan	nju	zju

Two commonly used rank correlation coefficients, Spearman's ρ and Kendall's τ, are employed to assess the accuracy of the computed node ranks with ResearchRank. Given two ranks R(X) and R(Y), Spearman's ρ is defined as:

$$\rho = \frac{\text{cov}(\text{R}(X), \text{R}(Y))}{\sigma_{\text{R}(X)}\sigma_{\text{R}(Y)}},$$

where cov$(\text{R}(X), \text{R}(Y))$ is covariance of the rank variables, $\sigma_{\text{R}(X)}$ and $\sigma_{\text{R}(Y)}$ are the standard deviations of the rank variables.
Let $(x_1, y_1), ..., (x_n, y_n)$ be a set of observations of the joint random variables X and Y, any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be concordant if the sort order of (x_i, x_j) and (y_i, y_j) agrees, otherwise they are said to be discordant. Kendall's τ is defined as:

$$\tau = \frac{\#\text{concordant pairs} - \#\text{discordant pairs}}{\#\text{pairs}}.$$

3 Node ranking methods

3.1 Node ranking based on centralities

DC, CC, and BC are three commonly used measures of node importance in network analysis. DC measures the number of edges that are incident to a node, indicating the degree to which a node is connected to other nodes in the network. The definition of normalized DC is as follows:

$$\text{DC}(i) = \frac{d_i}{n-1},$$

where d_i is the degree of the node i; and n is the total number of nodes in the network.
CC measures the average distance between a node and all other nodes in the network, reflecting the node's ability to rapidly communicate with other nodes. The normalized CC is defined as:

$$\text{CC}(i) = \left[\frac{\sum_{j=1}^{N} d(i,j)}{n-1}\right]^{-1},$$

where $d(i, j)$ is the length of the average shortest path between node i and j.
BC measures the extent to which a node lies on the shortest paths between other nodes in the network, indicating its potential to control the flow of information or resources between different parts of the network. The BC of node i is defined as:

$$\text{BC}(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}},$$

where g_{jk} is the number of geodesics connecting node j and k, and $g_{jk}(i)$ is the number that actor i is on.

3.2 PageRank

The widely used PageRank algorithm is based on the idea that a web page's importance is determined by the importance of the pages that link to it. The algorithm models people's visits to websites as a

Markov chain on a connected graph, denoted by $G = \{V, E, W\}$. The transition according the links between two websites is reflected by

$$P_1 = D^{-1}W, \quad D := \textbf{diag}\left(\sum_{j=1}^{|V|} w_{ij}\right).$$

The algorithm also models random visits to websites using the matrix $E = \frac{1}{|V|}\mathbf{1}\mathbf{1}^T$. By combining these two types of transitions, the final transition matrix is obtained as

$$P_\alpha = \alpha P_1 + (1-\alpha)E, \tag{1}$$

where α is a hyperparameter that determines the weight given to the importance of the links.

3.3 Weighted PageRank

In the original PageRank algorithm, all links between pages are considered to have equal weight. However, in many real-world networks, the links between nodes may have different strengths or weights. WPR takes this into account by assigning a weight to each link based on its importance. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as W_{ij}^{in} and W_{ij}^{out}, respectively.

W_{ij}^{in} is the weight of $link(i,j)$ calculated based on the number of inlinks of page i and the number of inlinks of all reference pages of page j.

$$W_{ij}^{in} = \frac{I_j}{\sum_{p \in R(i)} I_p}$$

where I_j and I_p represent the number of inlinks of page j and page p, respectively. $R(i)$ denotes the reference page list of page i.

Similarly, $W_{(ij)}^{out}$ is the weight of $link(i,j)$ calculated based on the number of outlinks of page j and the number of outlinks of all reference pages of page i.

$$W_{(ij)}^{out} = \frac{O_j}{\sum_{p \in R(i)} O_p}$$

where O_j and O_p represent the number of outlinks of page j and page p, respectively. $R(i)$ denotes the reference page list of page i.

Considering the importance of pages, the original PageRank formula is modified as

$$PR(j) = (1-\alpha) + \alpha \sum_{i \in B(j)} PR(i) W_{ij}^{in} W_{ij}^{out}$$

where $B(j)$ is the set of pages that point to j. $PR(i)$ and $PR(j)$ are rank scores of page i and j, respectively.

3.4 HITS Ranking

HITS is a popular algorithm for ranking web pages based on their authority and hub. An authority page is one that is linked to by many hub pages, while a hub page is one that links to many authority pages. The algorithm iteratively computes two scores for each page: the authority score and the hub score. Firstly, each webpage is assigned two scores, authority score x_i and hub score y_i. The mutually reinforcing relationship is represented as

$$x_i' = \sum_{e_{ji} \in E} y_j, y_i' = \sum_{e_{ij} \in E} x_j, x_i = x_i' / \|x_i'\|, y_i = y_i' / \|y_i'\| \tag{2}$$

where $\|\cdot\|$ is L_2 norm. Iteratively solving Eq. 2 and finally we will obtain stable solutions x_i^*, y_i^*. Write $L_{ij} = w_{ij}$ if $e_{ij} \in E$ and 0 otherwise. And then we obtain an adjacency matrix L. Then Eq. 2 can be written as

$$x' = L^T y, y' = Lx, x = x'/\|x\|, y = y'/\|y\|.$$

Let $x^{(t)}, y^{(t)}$ denote hub and authority score in t iteration. Then the iteration can be rewritten as

$$c_1 x^{(t+1)} = L^T L x^{(t)}, c_2 y^{(t+1)} = L L^T y^{(t)},$$

starting with $x^0 = y^0 = (1, 1, \ldots, 1)^T$, where $c_i, i = 1, 2$ are scaling parameter such that $\left\|x^{(t+1)}\right\| = \left\|y^{(t+1)}\right\| = 1$. The final solution x^*, y^* are the principal eigenvectors of $L^T L$ and LL^T, respectively. Thus we also can directly calculate the singular value decomposition of L to obtain the final solutions.

3.5 SALSA algorithm

SALSA combines the ideas of HITS and PageRank, assigning hub and authority values to web pages and using a Markov chain to calculate page rankings. Unlike PageRank, SALSA selects highly relevant web pages after receiving a user's query request and assigns pages directly linked to the root set as an “expand set". The algorithm then determines rankings based on link analysis within the expand set. After getting the page ranking, the algorithm divides the pages in the expand set into a bipartite graph consisting of a hub set and an authority set. The output from the hub set constitutes the edge of the bipartite network. Unlike HITS, SALSA does not adopt the authority-hub mutual enhancement method but instead adopts PageRank's random walk.

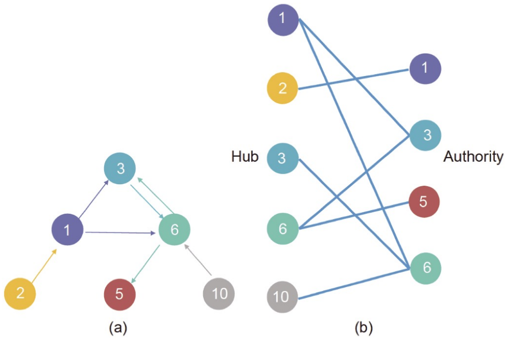

Figure 1: The origin graph (a), and the bipartite graph (b).

Fig. 1 present an example of dividing the set of nodes into a bipartite network. Denote V_h the hub set and V_a the authority set, we have

$$V_h = \{1, 2, 3, 6, 10\}, V_a = \{1, 3, 5, 6\}.$$

SALSA uses both row and column weighting to compute its hub and authority scores. Let L_r be be L with each nonzero row divided by its row sum and L_c be L with each nonzero column divided by its column sum. In this example,

$$L = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, L_r = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, L_c = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix}.$$

Then, the co-citation matrix A consists of the nonzero rows and columns of $L_c^T L_r$ and the bibliographic coupling matrix H consists of the nonzero rows and columns of $L_r L_c^T$. By calculation, we have

$$H = \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix}, A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{6} & 0 & \frac{5}{6} \end{pmatrix}$$

where the columns and rows of the above two matrices are corresponding to the hub set and authority set.

4 Results

4.1 PageRank with various α

Applying the PageRank method to Chinese universities' dataset, we can get a sequence of rankings PageRank$_\alpha$ with different α. We set $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.85, 0.9\}$ and compare these ranking results with ResearchRank by Spearman's ρ and Kendall's τ.

Table 2: Comparison of PageRank$_\alpha$ and ResearchRank.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.85	0.9
Spearman's ρ	0.672	0.674	0.681	0.686	0.691	0.699	0.700	0.706	**0.706**
Kendall's τ	0.489	0.493	0.498	0.503	0.507	0.512	0.516	0.520	**0.521**

As shown in Table 2, both Spearman's ρ and Kendall's τ values increase as α gets larger. Based on Eq. 1, when a higher damping factor α is used, the importance of the links is given more weight, and the scores of the pages that are linked to by other important pages tend to be higher. The results may indicate that web graph of Chinese universities has a strong structure, with many pages linking to a few highly important pages. As a consequence, PageRank can reflect the research level of a university.

Table 3: Top-5 universities of PageRank with different α.

α	1	2	3	4	5
0.1	tsinghua	pku	uestc	nju	sjtu
0.2	tsinghua	pku	uestc	nju	sjtu
0.3	tsinghua	pku	nju	sjtu	uestc
0.4	tsinghua	pku	sjtu	nju	uestc
0.5	tsinghua	pku	sjtu	nju	uestc
0.6	tsinghua	pku	sjtu	nju	uestc
0.7	tsinghua	pku	sjtu	nju	uestc
0.85	tsinghua	pku	sjtu	nju	uestc
0.9	tsinghua	pku	sjtu	nju	uestc

Table 3 shows the top-5 universities based on PageRank score with different α. "tsinghua" and "pku" maintain their positions as the top two universities with the highest PageRank scores across all damping factors. The relative order of "sjtu", "nju", and "uestc" varies with different α values. It is worth noting that "fudan" and "zju", which are among the top-5 universities based on ResearchRank, do not feature in the top-5 universities based on PageRank.

4.2 Performance of other node ranking methods

We apply the node ranking methods introduced in Section 3 to Chinese universities' dataset. Specifically, we utilize the PageRank (PR) and WPR algorithms, with a damping factor α of 0.9, as described in Section 4.1. The numerical results are summarized in Table 4.

Table 4: Comparison between different node ranking results and ResearchRank.

Method	DC	CC	BC	PR	WPR	HITS Hub	HITS Authority	SALSA Hub	SALSA Authority
Spearman's ρ	0.439	0.645	0.449	0.706	0.593	0.540	**0.750**	0.440	0.722
Kendall's τ	0.309	0.472	0.297	0.521	0.418	0.378	**0.572**	0.313	0.551

Among the three centrality-based ranking methods, CC has the highest Spearman's ρ and Kendall's τ with ResearchRank. These results suggest that nodes located in close proximity to other nodes in the network are likely to be important or influential. Additionally, the weblink network exhibits a high degree of interconnectedness and short path lengths between nodes.

In terms of authority rankings, the HITS algorithm was found to be most similar to ResearchRank, followed by SALSA and PageRank. On the other hand, the overall authority rankings were found to be more similar to ResearchRank compared to hub rankings. The results imply that the quantity of incoming links to a node is more important in determining the research rank of a university.

Table 5: Top-5 universities of different node ranking methods.

Method	1	2	3	4	5
DC	pku	tsinghua	nju	zsu	sjtu
CC	pku	tsinghua	nju	uestc	sjtu
BC	pku	tsinghua	sdu	bfsu	sjtu
PR	tsinghua	pku	sjtu	nju	uestc
WPR	tsinghua	pku	sjtu	zsu	whu
HITS Hub	pku	ustc	zsu	sjtu	zju
HITS Authority	tsinghua	pku	uestc	sjtu	nju
SALSA Hub	pku	ustc	zsu	njau	sjtu
SALSA Authority	tsinghua	pku	uestc	sjtu	nju

The top-5 universities identified by different node ranking methods are listed in Table 5. Notably, several universities such as "tsinghua", "pku", "nju", and "sjtu" consistently appear on most of the top lists across the ranking methods. These findings suggest that these universities are highly influential or important in the network, as they are consistently ranked highly by different methods.

Comparing the results of PR and WPR, the top-3 universities are the same and the fourth and fifth ranks of PR are "nju" and "uestc", while for WPR, they are "zsu" and "whu". It is because that the total indegrees of "nju" (340) and "uestc" (428) are higher than those of "zsu" (311) and "whu" (230). However, the total outdegrees of "zsu" (861) and "whu"(485) are higher than that of "nju" (270) and "uestc" (9). PR algorithm considers all links to have equal weights, which may not accurately reflect the true importance of each link. In contrast, the WPR algorithm assigns different weights to different links based on their types or sources, which may provide a more fair view of the node importance.

Comparing the results of HITS and SALSA algorithms, it was found that the authority rankings are exactly the same. However, the hub rankings differ for the fourth and fifth ranks. The total outdegree of "zju" and "sjtu" are 383 and 647, respectively, while that of "njau" is up to 688. Thus, the actual ranking of the webpage of "njau" is more likely to be higher than "zju", whose outdegree is only 383. However, "zju" ranks high in the HITS algorithm due to its high number of hyperlinks to "pku", which has the highest authority rank in the dataset. This is a weakness of the HITS algorithm, as it can be manipulated by webpages that use many hyperlinks to point to high-ranking webpages, leading to artificially inflated hub rankings.

5 Conclusion

This study aimed to compare different node ranking methods on the Chinese University Weblink dataset. The results showed that, for the PR algorithm, both Spearman's ρ and Kendall's τ increased as the damping factor α increased, indicating that the page rank tended to align more closely with ResearchRank. Additionally, the HITS authority rank had the highest correlation with ResearchRank compared to all other methods. Furthermore, the results highlighted the influence of assigning different weights in node ranking. These findings underscore the importance of carefully considering the underlying assumptions and limitations of different ranking methods when interpreting the results of network analysis.

References

[1] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, 2(1):113–120, 1972.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[3] Linton C Freeman et al. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263, 2002.

[4] Byungseok Kang, Daecheon Kim, and Hyunseung Choo. Internet of everything: A large-scale autonomic iot gateway. *IEEE Transactions on Multi-Scale Computing Systems*, 3(3):206–214, 2017.

[5] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[6] Ronny Lempel and Shlomo Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks*, 33(1-6):387–401, 2000.

[7] JiaQi Liu, XueRong Li, and JiChang Dong. A survey on network node ranking algorithms: Representative methods, extensions, and applications. *Science China Technological Sciences*, 64(3):451–461, 2021.

[8] Wenpu Xing and Ali Ghorbani. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE, 2004.