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

HU, Mingyun

Department of Mathematics

May 8, 2024

Table of Contents

Table of Contents

Introduction

Node ranking is a crucial problem in network analysis. It aims to measures the importance of nodes in a graph based on certain criteria.

Node ranking methods can be classified into three categories:

- Centrality-based ranking
  e.g., degree centrality, closeness centrality, betweenness centrality
- PageRank-based ranking
  e.g., PageRank, Weighted PageRank
- Hyperlink-induced topic search (HITS)-based ranking
  e.g., HITS algorithm, Stochastic approach for link structure analysis (SALSA)

Introduction

Identifying the most appropriate node ranking method for a given network and task may not be trivial, as different methods may capture different aspects of node importance.

The ranking of web link network can provide valuable insights into the web page's popularity and authority. In the case of university web link, the node rank may serve as an indicator of their research rank.

In this study, we explore the performance of different node ranking methods on the Chinese University Weblink dataset.

Table of Contents

Data and evaluation metrics

- The dataset contains 76 universities from mainland China.
- $W \in \Re^{76\times 76}$ is the link matrix whose (i,j)-th element gives the number of links from university i to j.
- ResearchRank denotes the universities' research ranking.

Table: Universities with top-5 ResearchRank.

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

Data and evaluation metrics

Two rank correlation coefficients are employed to assess the similarity of the computed node ranks with ResearchRank.

- Spearman's ρ:

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

  where $\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))$ is covariance of the rank variables, $\sigma_{\mathrm{R}(X)}$ and $\sigma_{\mathrm{R}(Y)}$ are the standard deviations of the rank variables.

- Kendall's τ:

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

Table of Contents

Centrality-based ranking

Degree centrality (DC), closeness centrality (CC), and betweenness centrality (BC) are commonly used centrality-based ranking measures.

- DC measures the number of edges that are incident to a node:

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

- CC measures the average distance between a node and all other nodes in the network:

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

- BC measures the extent to which a node lies on the shortest paths between other nodes:

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

PageRank

PageRank (PR) 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

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

The random visits to websites are modeled by the matrix $E = \frac{1}{|V|}\mathbf{1}\mathbf{1}^T$.

The final transition matrix is obtained as

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

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

Weighted PageRank

The original PR algorithm models all links to have equal weight.

Weighted PageRank (WPR) assigns weight to links based on the number of inlinks and outlinks of the page they are on.

The weights of the incoming and outgoing links of $link(i, j)$ are defined as

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

where I_k and O_k represent the number of inlinks and outlinks of page k, respectively. $R(i)$ denotes the reference page list of page i.

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

HITS algorithm

HITS algorithm ranks web pages based on their authority and hub scores. High in-degree webpages are regarded as *authorities*, high out-degree webpages are regarded as *hubs*.

We use primary **right singular vector** of W as scores to give the HITS-authority ranking, and primary **left singular vector** of W as scores to give the HITS-hub ranking.

The final scores are the principal eigenvectors of $W^T W$ and WW^T, respectively.

SALSA algorithm

SALSA combines the ideas of HITS and PageRank.

Like HITS, SALSA algorithm assigns a hub score and an authority score to each web page, and works on a focused bipartite graph.

Like PageRank, SALSA computes the scores by simulating a random walk through a Markov chain that represents the graph of web pages.

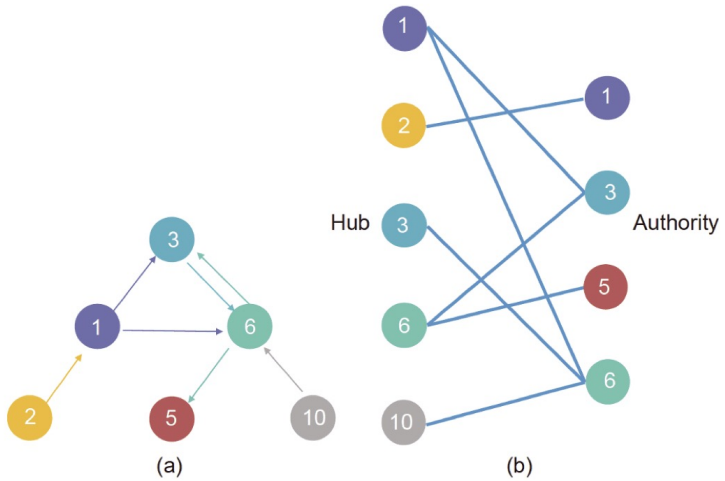

Figure: The origin graph (a), and the bipartite graph (b).

Table of Contents

PageRank with different α

Table: Comparison of ResearchRank and PageRank with different α.

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**

- Both Spearman's ρ and Kendall's τ values increase as α gets larger.
- The node ranking of universities' web links can reflect the research level of the institution.

PageRank with different α

Table: 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

- Tsinghua and pku remain the top two universities across all α.
- The relative order of sjtu, nju, and uestc varies with different α values.
- However, fudan and zju, which are among the top-5 universities of ResearchRank, do not appear on the list.

Performance of node ranking methods

Table: 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 centrality-based ranking methods, CC has the highest Spearman's ρ and Kendall's τ.
- The HITS-authority rank is the most similar to ResearchRank, followed by SALSA-authority rank and PageRank.
- The results imply that **the quantity of incoming links** to a web page is more important in determining the research rank of a university.

Performance of node ranking methods

Table: 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

- Tsinghua, pku, nju, and sjtu consistently appear on most of the top lists across the ranking methods.
- These universities are highly influential and important in the network.

Performance of node ranking methods

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

Method	1	2	3	4	5
PR	tsinghua	pku	sjtu	nju	uestc
WPR	tsinghua	pku	sjtu	zsu	whu

The fourth and fifth ranks of PR and WPR are different.
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 much higher than that of nju (270) and uestc (9).
PR algorithm considers all links to have equal weights. In contrast, the WPR algorithm assigns different weights to different links based on their sources, which may provide a more fair view of the node importance.

Performance of node ranking methods

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

Method	1	2	3	4	5
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 hub ranks differ for the fourth and fifth ranks.
The total outdegree of zju, sjtu, and njau are 383, 647, and 688. The actual ranking of the webpage of njau is more likely to be higher than zju. However, zju ranks high in the HITS algorithm due to its high number of hyperlinks to pku, which has the highest authority rank.
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.

Table of Contents

Conclusion

- For the PR algorithm, both Spearman's ρ and Kendall's τ increase as the hyperparameter α increases.
- The HITS-authority rank has the highest correlation with ResearchRank compared to all the other methods.
- The results demonstrate the influence of assigning different weights in node ranking.

Thank You!