CSIC 5011 Mini-Project 2: Analysis of Chinese (mainland) University Weblink Data

Qi Bu, qbu@connect.ust.hk
Department of Physics, HKUST

Introduction

The data set contains 76 universities and the element *W(i, j)* of the link matrix *W* stands for the number of links from university *i* to *j*. I would like to first make use of the PageRank and HITS ranking to do comparison and further using MDS and diffusion distances to test whether there are clusters among the 76 universities.

PageRank & HITS

From Figure 1 we can see the PageRank values change with the parameter α. As α is small, the network is close to a totally random search, which means in every time step the "random walker" may randomly choose the next webpage among all 76 universities with equal probability. However, when α increases, the difference in PageRank Values between universities becomes larger. The top 2 universities "pku.edu.cn" and "tsinghua.edu.cn" stand out from others.

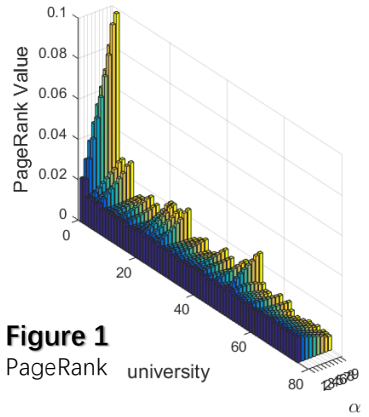
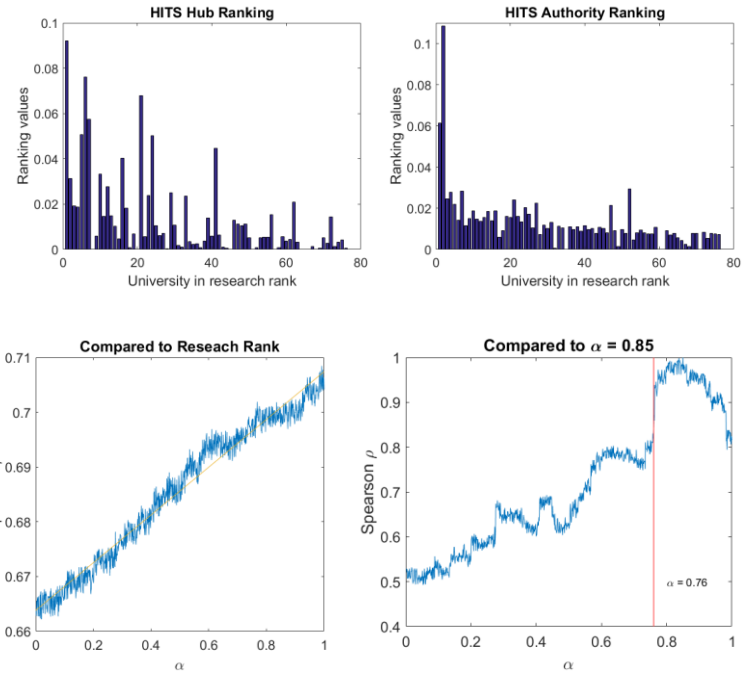


Figure 1 PageRank

However, from the figure, we may also find that the PageRank value changes monotonically with α.

On the figure in the middle I plot the correlation coefficient of ranking compared with Google's choice ($\alpha = 0.85$) vs. α.

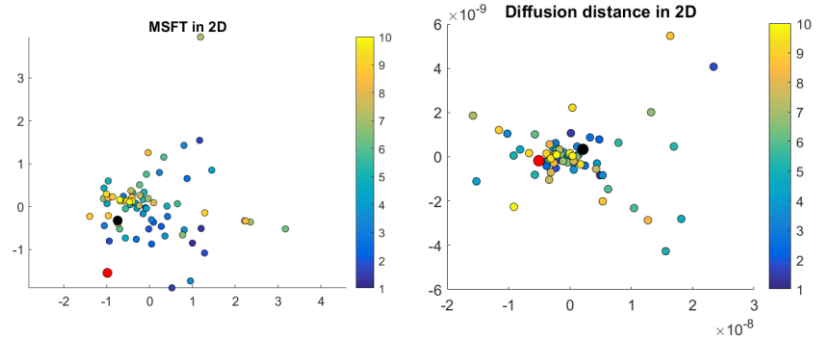
HITS Ranking values



Random Walk

Since random walk is assumed in PageRank, we may consider using other concepts of random walk, for example mean first passage time. I would like to compute the mean first passage time between each two universities and use the average of MSFT from *i* to *j* and from *j* to *i* as the distance between *i* and *j*. Then MDS can be applied. The same strategy can also be used to diffusion distances.

Spearman's ρ	Research Rank	PageRank	HITS Hub	HITS Authority
Research Rank	**1**	**0.4386**	**0.5422**	**0.7505**
PageRank		**1**	**0.2453**	**0.3924**
HITS Hub			**1**	**0.4459**
HITS Authority				**1**



Analysis on MDS Results

For random walk, we assume that during each time step, the random walker will select one of the out-linked neighbors of the current website to move to, and the probability of moving is proportional to the number of links from this node to its neighbors.
On the two MDS figures above, each dot represents a university and the color of the dots is the research rank. The big red dot is 'pku.edu.cn' and the black dot is 'tsinghua.edu.cn'. On the left figure we can find a green dot far away from other dots, which is 'uestc.edu.cn'. I found that this website has many inlinks but few outlinks. Maybe that makes a random walker hard to start from this website and reach other low-rank websites. The same property can be found on the upper-right two nodes(yellow and blue) of the figure using diffusion distances.