Project 3

Name: Caleb Thian Jia Le 田家樂

Student ID: NN6114035

Find a Way

The main idea is to increase the link on node 1, but not increases the too much on others, and because need to increase both hub and authority, thus an edge link into node 1 and an edge link out from node 1 is necessary.

Ways:

1. Add edge out from node 1 (ex. 1→4)

2. Add edge link to node 1 (ex. 3→1)

Note that the 2 edges add should not be occur on the same node.

	Before			After		
	Hub	Authority	PageRank	Hub	Authority	PageRank
graph_1.txt	0.2	0	0.025	0.5	0.25	0.101
graph_2.txt	0.2	0.2	0.2	0.357	0.357	0.278
graph_3.txt	0.191	0.191	0.173	0.322	0.262	0.224

The below shows the graph after process, the graph on the left is the original and the graph on the right is the graph after adding edges:

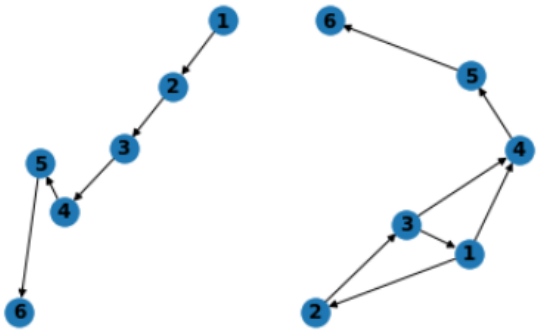

Figure 1：Graph 1 before and after process

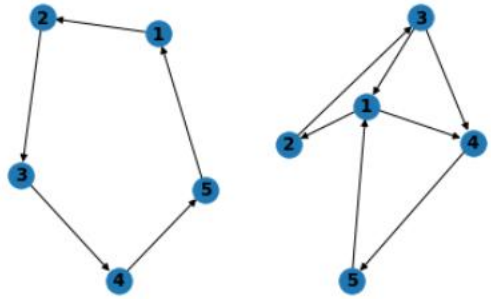

Figure 2: Graph 2 before and after process

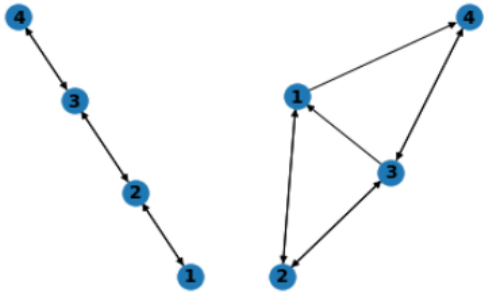

Figure 3: Graph 3 before and after process

Algorithm Description

PageRank

1. Initialize PageRank of each node with $\frac{1}{n}$, where n is the number of nodes.
2. Update PageRank of each node by the formula $PR(v_i) = \frac{d}{n} + (1 - d) \sum_{e_{j,i} \in E} \frac{PR(v_j)}{Outdegree(v_j)}$, where d is the damping factor, $e_{i,j}$ is an edge that link to v_j from v_i, $PR(v_i)$ is the PageRank of v_i, and $Outdegree(v_i)$ is the out degree of v_i. The damping factor is set to 0.1.
3. Repeat Step 2 until the PageRank converges.

HITS

1. Initialize authority and hub of each node with 1.
2. For each node,
   a) update authority by the sum of hub of its parents.
   b) update hub the sum of authority of its children.
3. Normalize the authority and hub.
4. Repeat Step 2 and Step 3 until converges.

SimRank

1. Initialize SimRank as $n \times n$ identity matrix. This indicates that $S(a,b) = \begin{cases} 1, if\ a = b \\ 0, otherwise \end{cases}$.
2. Update the SimRank matrix by formula

$$S(a,b) = \begin{cases} 1, if\ i = j \\ \dfrac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S(I_i(a), I_j(b)), otherwise \end{cases}$$

   where C is the decay factor, set as 0.7, and I(a) represent the parents of a.
3. Repeat Step 2 by 30 iterations.

Result Analysis and Discussion

PageRank

PageRank is a way of measuring the importance of a node. To simplify the analysis, only the maximum value and corresponding node will be recorded. The table below shows the results under different damping factor

Graph	d	maximum value	corresponding node
graph_1.txt	0.1	0.379	6
	0.125	0.353	6
	0.15	0.332	6
graph_2.txt	0.1	0.2	all
	0.125	0.2	all
	0.15	0.2	all
graph_3.txt	0.1	0.327	2,3
	0.125	0.326	2,3
	0.15	0.324	2,3

Analysis and discussion:

1. According to the result of graph 1, the closer the node to the leaf, the more important the node is.
2. According to the result of graph 2, the PageRank of the nodes in a circle will be equal, which is interpretable.
3. According to the result of graph 3, the more the link of a node, the more important the node is.
4. According to the result of graph 1 and graph 3, with increasing the damping factor, the range of the PageRank will be decreased. Thus, PageRank between nodes will be more balance.

HITS

HITS is the way to measure the hubs and authorities of a graph. The idea of hubs and authorities can be used to rank the importance of nodes in a graph or network. To simplify the analysis, only the maximum value and the corresponding node will be recorded. The table below shows the results:

Graph	Hub/Authority	maximum value	corresponding node
graph_1.txt	Hub	0.2	all except 6
	Authority	0.2	all except 1
graph_1.txt	Hub	0.2	all
	Authority	0.2	all
graph_1.txt	Hub	0.309	2,3
	Authority	0.309	2,3

Analysis and discussion:

1. According to the result of graph 1, the value of hub and authority are respect to the indegree and outdegree of the node. Similar to the result in graph 2 and graph 3

SimRank

SimRank is a measure of the similarity of two objects in a graph or network. It is based on the idea that two objects are similar if they are related to similar parents in a directed graph. To simplify the analysis, only the maximum value (except 1 as 1 occur during the

common nodes pair) and the corresponding nodes pair will be recorded. The table below shows the results under different decay factor:

Graph	C	maximum value except 1	corresponding nodes pair
graph_1.txt	0.7	0	(a,b) where a and b are distinct nodes
	0.8	0	(a,b) where a and b are distinct nodes
	0.9	0	(a,b) where a and b are distinct nodes
graph_2.txt	0.7	0	(a,b) where a and b are distinct nodes
	0.8	0	(a,b) where a and b are distinct nodes
	0.9	0	(a,b) where a and b are distinct nodes
graph_3.txt	0.7	0.538	(1,3) and (2,4)
	0.8	0.667	(1,3) and (2,4)
	0.9	0.818	(1,3) and (2,4)

Analysis and discussion:

1. According to the result of graph 1 and graph 2, there is not similar node pairs exists in these two graphs as every node are unique, there is no exists two nodes with similar parents.
2. According to the result of graph 3, node 1 is similar to node 3 as they have common parent node 2, similar for node 2 and node 4.
3. According to the result of graph 3, by increasing the decay factor, the SimRank of the similar pairs will be increased. It indicates that the decay factor controls the value of the SimRank.

Effectiveness Analysis

The table below shows the execution time in seconds for each graph under different methods, note that the damping factor = 0.1, decay factor = 0.7, and maximum iteration = 30.

	number of nodes, N	number of edges, E	PageRank	HITS	SimRank
graph_1.txt	6	5	0	0	0.01
graph_2.txt	5	5	0	0	0.00
graph_3.txt	4	6	0	0	0
graph_4.txt	7	18	0	0	0.03
graph_5.txt	469	1102	0.03	0.02	12853.73
graph_6.txt	1228	5220	0.06	0.29	
ibm-5000.txt	836	4798	0.03	0.05	

Assume that PageRank and HITS will converges in 30 loops, then the time complexity of PageRank will be O(N + E), which mainly dominated in step 2, as every edge will be considered only once in each loop, and the loop iterates over all the nodes. The time complexity of HITS will be O(N+2E) = O(N+E) too, which mainly dominated in step 2, as each edge states a parent and a children, each will be considered once in each loop of step 2 and 3, and the loop iterates over all the nodes too. In SimRank, the loop mainly iterates over all the combination of nodes, and the maximum iteration is a constant, thus the time complexity will be $O(N^2)$.

According to the time complexity analysis above, PageRank and HITS can be done really fast, in fact, these algorithms finished in less than 1 seconds over all the graphs given. But for SimRank, it finished slower, especially when it met graph 5, which has 469 nodes, it needed about 3.7 hours to finish the iterations. It is slower in graph 5 because of the number of nodes(N), forming N^2 pairs of combination to be iterated.