**Nick Petty**

Z23296080

**Homework 2 Node Similarity and Community Detection**

**Question 1 [1.5 pts]** Please use your own language to briefly explain the following concepts:

PageRank score: a way of determining the importance of a web page by counting the number and quality of links to that page.

Rooted PageRank: a modification to PageRank that keeps the random walker focused on a particular area of the graph. This is done by randomly resetting the walker back to the root node being investigated.

Network community: a group of nodes in a network which are densely connected to each other compared to their connections with the rest of the network.

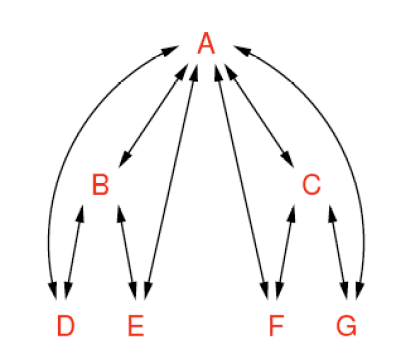
Clique: a subset of nodes in a network that forms a complete subgraph. Each node in the group is connected to all other nodes in the group.

k-Clique: a subgraph where the shortest path between any two nodes is <= k, and which cannot maintain this feature if more nodes are added.

Low-rank approximation: by computing the rank-k matrix of a graph, choosing a low value of k, and using another community detection method, the noise in a graph can be reduced. This is also used in conjunction with Singular Value Decomposition to find significant data amid noise.

**Question 2 [2 pts]:** Given seven web pages with the following link structure,

1. Please use “Power Iteration” (a.k.a simple iteration) to calculate the PageRank scores for each website. (You only need to show the first and the second iterations results, with the initial PageRank scores for each node being set as 1/n=0.15) [1 pt].
2. Please also use Eigenvector based approach to calculate PageRank scores for each web page [1 pt] (please show your solutions.)



1. The adjacency matrix is normalized and the PageRank Power Iteration is applied.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | G |  |  | A | B | C | D | E | F | G |
| A | 0 | 1 | 1 | 1 | 1 | 1 | 1 |  | A | 0.0000 | 0.3333 | 0.3333 | 0.5000 | 0.5000 | 0.5000 | 0.5000 |
| B | 1 | 0 | 0 | 1 | 1 | 0 | 0 |  | B | 0.1667 | 0.0000 | 0.0000 | 0.5000 | 0.5000 | 0.0000 | 0.0000 |
| C | 1 | 0 | 0 | 0 | 0 | 1 | 1 |  | C | 0.1667 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.5000 | 0.5000 |
| D | 1 | 1 | 0 | 0 | 0 | 0 | 0 |  | D | 0.1667 | 0.3333 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| E | 1 | 1 | 0 | 0 | 0 | 0 | 0 |  | E | 0.1667 | 0.3333 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| F | 1 | 0 | 1 | 0 | 0 | 0 | 0 |  | F | 0.1667 | 0.0000 | 0.3333 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| G | 1 | 0 | 1 | 0 | 0 | 0 | 0 |  | G | 0.1667 | 0.0000 | 0.3333 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 |
| A | 0.1500 | 0.4000 | 0.2667 | 0.3444 | 0.2963 | 0.3272 | 0.3070 |
| B | 0.1500 | 0.1750 | 0.1417 | 0.1694 | 0.1491 | 0.1633 | 0.1536 |
| C | 0.1500 | 0.1750 | 0.1417 | 0.1694 | 0.1491 | 0.1633 | 0.1536 |
| D | 0.1500 | 0.0750 | 0.1250 | 0.0917 | 0.1139 | 0.0991 | 0.1090 |
| E | 0.1500 | 0.0750 | 0.1250 | 0.0917 | 0.1139 | 0.0991 | 0.1090 |
| F | 0.1500 | 0.0750 | 0.1250 | 0.0917 | 0.1139 | 0.0991 | 0.1090 |
| G | 0.1500 | 0.0750 | 0.1250 | 0.0917 | 0.1139 | 0.0991 | 0.1090 |

2. The eigenvalues and eigenvector were calculated with bluebit.gr, with the highest value highlighted to indicate the corresponding eigenvector. The first table is from the adjacency matrix; the second table is from the normalized adjacency matrix.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Eigenvalues | Eigenvector |  | Eigenvalues | Eigenvector |
| 3.2361 | 0.6015 |  | 1.0000 | 0.7171 |
| 2.0000 | 0.3717 |  | 0.3333 | 0.3586 |
| 1.2361 | 0.3717 |  | 0.6667 | 0.3586 |
| 1.4142 | 0.3008 |  | 0.5773 | 0.2391 |
| 1.4142 | 0.3008 |  | 0.5773 | 0.2391 |
| 0.0000 | 0.3008 |  | 0.0000 | 0.2391 |
| 0.0000 | 0.3008 |  | 0.0000 | 0.2391 |

In both methods, the nodes are ranked as follows:

Rank 1 – A

Rank 2 – B, C

Rank 3 – D, E, F, G

**Question 3 [1 pt]:** In Quesiton 2, please use rooted PageRank to calculate similarity between each pair of nodes. Each time, the random walker has a probability 1-*α*  (where *α=0.2*)to return back toan original node*.* (Please show your solutions).

**Question 4 [1.5 pts]:** The following networks show connections between 8 individuals in a small community. For node pairs (1, 7) and (1, 6), please use following measures to calculate their similarity (or distance) value and conclude which pair is more likely to form a link.

* 1. Jacarrd’s Coefficient (0.25 pt)
  2. Adamic/Adar (0.25 pt)
  3. Preferential attachment (0.25 pt)
  4. Katz (with β=0.05) (0.25 pt)
  5. SimRank score with C=1 (please show the SimRank score after the 1st iteration). (0.5 pt)

1

2

4

5

6

7

8

3

**Question 5 [4 pts]:** In the following network,

1. Please find the complete set of communities by using 3-clquie [0.25pt], 3-club [0.25pt], and 3-core [0.25pt], respectively (If there are multiple sets, please just report the top three sets with the maximum number of nodes).
2. Please calculate the Geodesic distance between each pair of nodes, and use Multidimensional Scaling (MDS) to convert the network into a two dimensional space. Please report the values of all nodes in the two dimensional space and draw all nodes in the two dimensional space [1.25 pt].
3. Implement a k-means clustering algorithm (selecting k=2 and using node 18 and node 1 as the initial centers), and report the community structures after 10 iterations (You may use any other third party tools for k-means clustering. Or you can follow the k-means Excel implementation in the following URL to calculate the results) [2 pts]

k-means: <http://www.csse.monash.edu.au/courseware/cse5230/2004/assets/clustering.pdf>

3

1

2

4

5

6

7

13

17

18

16

9

14

11

8

12

10

15