TSKS11 Hands-On Session 2

Fall 2019

We will use real data consisting of a subset of the English Wikipedia project. Nodes in the network correspond to Wikipedia pages, and edges in the network correspond to links between the pages. There are twenty sub-networks available, numbered by $N = 1, 2, 3, \ldots, 20$. The files `titles/`N`.txt` and `links/`N`.txt` contain the Nth sub-network.

- The file `titles/`N`.txt` contains the titles of the page/article that the nodes represent. The title of node i is found on row i in the file.
- A file `links/`N`.txt`, that contains all the links between the nodes. Each row in the file represents a link from the node in the left column to the node in the right column. Note that the network is directed.

To get started, randomly select (at least) five of the 20 available networks.

Note: built-in functions for centrality in SNAP, NetworkX, or similar libraries may **not** be used to solve tasks 2–5.

Task 1

Calculate the in and out degrees of all articles. Show the results in two lists; one list shows the in and out degrees of the five nodes with highest in-degree; one list shows the in and out degrees of the five nodes with highest out-degree.

Task 2

Calculate the hub and authority centrality of all articles. Display the result in the same way as in the previous task.

Task 3

Calculate the Katz Centrality of each article with

$$\alpha = 0.85 \cdot \frac{1}{|\lambda_{\text{max}}|},$$

where λ_{max} is the largest eigenvalue of the adjacency matrix. List the top five articles and their Katz score.

Task 4

Calculate the Google PageRank score of each article in your network with parameter $\alpha = 0.85$. List the top five articles and their PageRank score. Use the version of PageRank explained in the lecture notes. Also try some other values of α and comment on the result.

Task 5

Implement the iterative version of PageRank. Compare the result after one, two, five, ten and 100 iterations with the "exact" solution to the equation system obtained in Task 4. Show how the top 5 nodes evolve over the iterations and show how close the iterative solution is to the exact. (The latter is most easily shown by looking at the norm of the difference of the exact ranking and the iterative solution.)

Task 6

Comment on the results of Tasks 4 and 5. Considering the top 5 articles, and in particular the "winner", is the result plausible?

Reflect on similarities/differences between the result of the different centrality metrics.

Hints

- **Importing data to Matlab.** The following Matlab code can be used to import the data files (`links/1.txt` and `titles/1.txt` in the example):

```
filenumber = 1;
links = load(sprintf('/courses/tsks11/ht2019/data_and_fcns/session2/links/%d.txt',filenumber));
tit_file = fopen(sprintf('/courses/tsks11/ht2019/data_and_fcns/session2/titles/%d.txt',filenumber));
titles = textscan(tit_file,'%s');
titles = titles{1};
```

Now `links` is an $L \times 2$ edge list, and `titles` is an $N \times 1$ cell containing the titles, where L is the number of edges and N is the number of nodes (Wikipedia articles) in this network. Of course, replace the value of the variable `filenumber` to import other files (1 to 20).

- **Print the titles in Matlab.** Since the titles are now contained in a cell structure, one must use curly brackets { } to access the content (as opposed to vectors and matrices where we use parentheses (), also called round brackets). Suppose we want to print the titles corresponding to node indices 2, 4, 6, 9 and 10. One way to do that is this:

```
for i = [2 4 6 9 10]
fprintf(’%s\n’, titles{i})
end
```

Examination

- The program code you have written should be submitted to Urkund: `ollab13.liu@analys.urkund.se`.
- Collaboration on this homework in small groups is encouraged, but each student should perform programming work individually, and individually demonstrate understanding of all tasks.
- Library functions from SNAP, Matlab and the Python standard library may be used freely, but copying of code from other libraries and/or toolboxes, from the Internet, from other students, or from previous years’ students is prohibited.
- Individual oral examination takes place in class (computer lab).