Università degli Studi di Padova

Department of Information Engineering

Master degree in ICT for Internet and Multimedia

Course: Network Science

Homework 2

Davide Carta

ID: 1210702

January 28^{th} 2020

This homework aims at extracting relevant analytics from a network. For this task MATLAB was used.

About the dataset

The dataset contains data collected about Facebook pages (November 2017) and represents blue verified Facebook pages network of 13866 athletes. Nodes represent the pages and edges are mutual likes among them. This dataset was taken from the GitHub repository https://github.com/benedekrozemberczki/datasets/tree/master/facebook_page_page/sport/ .

Network overview

The network was made undirected considering that when a page likes another one, then a connection between them exists.

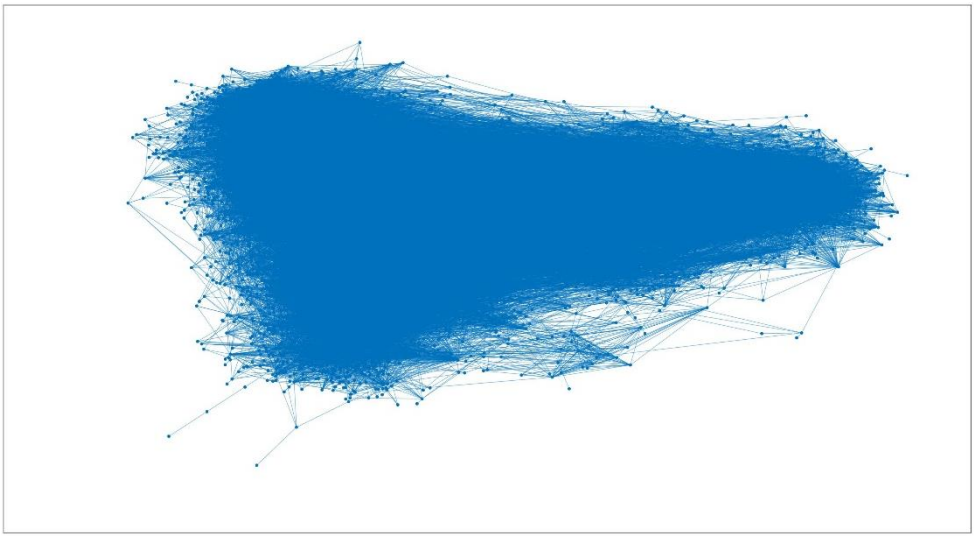

Figure 1. *Undirected network graph*

1. Ranking measures

To perform ranking measures, *PageRank* and *HITS* approaches were used.

1.1 PageRank

From the theory is known that the implementation of the *PageRank* algorithm is performed by mean of the power iteration method that follows the relation:

$$p_{t+1} = cMp_t + (1\text{-}c)q$$

with:

- **c** ≙ damping factor (set to 0.85)
- **M** ≙ normalized adjacency matrix
- **q** ≙ teleportation vector

id	name	new_id	r
1.7947e+14	{'Karolina Winkowska'}	264	0.0021591
1.5054e+15	{'Quincy Promes'}	9686	0.0021567
1.6383e+14	{'Anquan Boldin'}	878	0.0014649
5.811e+14	{'Lukas Verzbicas'}	8138	0.0013896
2.8693e+14	{'Mike Lorenzo-Vera'}	2845	0.001377
1.7557e+11	{'EC VSV'}	3976	0.0012629
4.3032e+14	{'Tony Romo'}	6505	0.0012358
1.3746e+14	{'Michigan Men's Golf'}	9733	0.0012317
1.1042e+11	{'TJ Schiller'}	3843	0.0011553
1.302e+11	{'Minnesota Women's Track & Field and Cross Country'}	4439	0.0011039

Table 1.1. *PageRank scores*

1.2 HITS

The HITS (*Hyperlink Induced Topic Search*) algorithm finds the hubs and authorities by mean of the following relations:

$$A = Ah$$

$$h = A^T a$$

with:

- **a** ≙ authorities score
- **h** ≙ hubs score

id	name	new_id	p
1.2681e+15	{'Abdoulaye DoucourÃ©' }	8924	-7.9457e-10
3.7859e+14	{'Mitch Whitmore' }	12543	-4.8838e-09
1.5138e+11	{'Bryan Clay' }	3204	-4.8838e-09
4.4325e+14	{'Craig McMorris' }	3962	-4.8838e-09
1.2015e+14	{'University of Delaware Fightin' Blue Hens'}	10230	-4.8838e-09
4.8366e+14	{'Andrew Suniula' }	3040	-1.2453e-08
9.5686e+14	{'JÃ©rÃ©my De Sousa' }	1836	-1.7571e-08
2.2686e+14	{'Alex Sorgente' }	10428	-2.1127e-08
3.6913e+14	{'Ã‰ngela Corina Clavijo' }	819	-2.2723e-08
6.1399e+14	{'Alfredo Talavera DÃ-az' }	3865	-2.2925e-08

***Table 1.2**. Hits scores*

From theory is known that the same result would appear for the authorities since we are dealing with an undirected network. For this reason, the result of the comparison between the two approaches is shown just for the hubs.

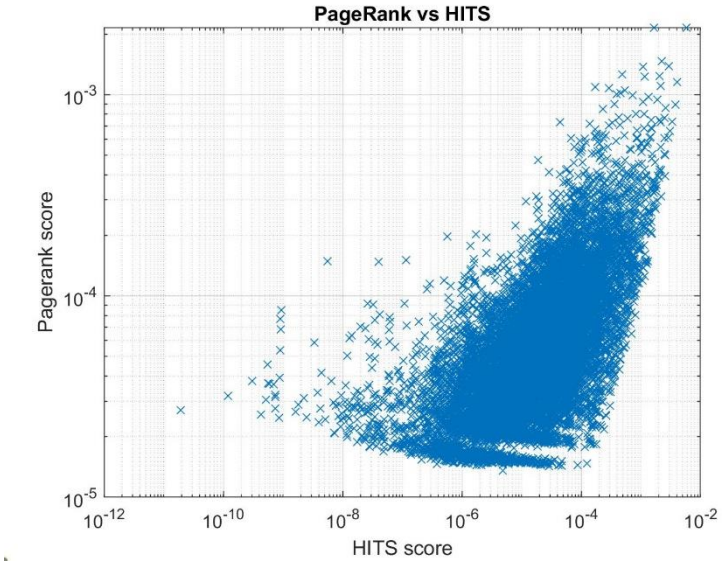

***Figure 2**. PageRank vs Hits scores*

2. Similarity measures

For this task the *SimRank* algorithm was used. SimRank is a general similarity measure that works according to the rule "two objects are considered to be similar if they are referenced by similar objects". The base case is that objects are 100% similar to themselves. This algorithm works based on the PageRank algorithm, with a different teleportation vector, having all zero components except of the one related to the position of the page that we are interest in. A particular page was chosen and the SimRank algorithm performed to find its top-10 similar ones.

The selected page is Mirza Teletovic, a former bosnian basketball player. (*https://it.wikipedia.org/wiki/Mirza_Teletović*)

id	name	new_id	sim_r
2.2476e+14	{'Westchester Knicks' }	13251	0.03754
2.6426e+10	{'Brooklyn Nets' }	4611	0.036179
4.8504e+14	{'Canyon Factory Racing' }	9785	0.036125
4.5898e+10	{'Phoenix Suns' }	3838	0.035367
1.7597e+14	{'Leicestershire County Cricket Club'}	6221	0.011459
3.5802e+14	{'Kilian Pagliuca' }	12393	0.0059221
8.6229e+14	{'Jamia Fields' }	827	0.0055555
1.0266e+14	{'Lidl Starligue' }	7708	0.0049417
1.1559e+14	{'Sorana Cirstea' }	12663	0.0043882
3.5402e+11	{'University of Utah Women's Soccer' }	5522	0.0043396

***Table 2**. Top-10 pages related to Mirza Teletovic*

Among the top-10 similar pages can be seen the ones of “Brooklyn Nets” and “Phoenix Suns”, two out of the three last teams he has played before retiring in the NBA, most famous basketball championship of the United States.

3. Community detection

For this task the *Spectral Clustering* and the *PageRank Nibble* approaches were used.

3.1 Spectral clustering

The Spectral Clustering algorithm, partitions nodes into two sets based on the second smallest eigenvalue of the normalized Laplacian matrix defined as $L = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ (with **D** ≙ degree matrix).

The community membership corresponds to the signs of Fiedler's eigenvector V_{N-1}. The spectral clustering algorithm gets the dataset partitioned in two main communities of size 33 and 13833 respectively.

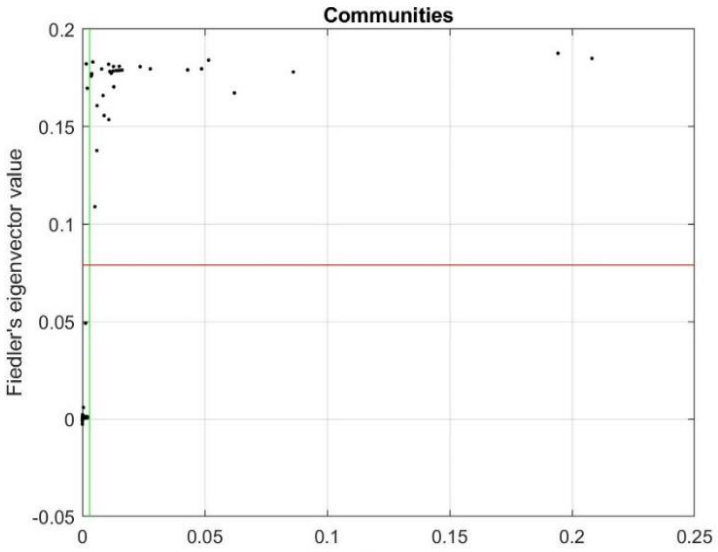

***Figure 3**.Spectral clustering community detection result*

3.2 PageRank Nibble

The PageRank Nibble algorithm is based on PageRank. It's initialized with some seed nodes to be used for the clusterization. A seed node *s* is chosen, and the PageRank Algorithm is ran, with teleportation vector **S** = {**s**}, meaning a whole zero vector with only one element set to one in the position of the chosen seed node. The idea behind this approach is that if 's' belongs to a good cluster, the random walk will get trapped inside it. This algorithm gets the dataset partitioned in two communities S and S^c of size 31 and 13835 respectively.

Due to the super connected network we are dealing with, no other significant partitions were found. For this reason, the algorithm stops after finding the first main one. The two approaches achieve an almost identical result, as can be seen by inspecting the smallest community obtained from the two approaches. To measure the quality of the cluster, the *minimum conductance* value is calculated as:

$$\varphi(s) = \frac{cut(S,S^c)}{\min\left(assoc(S),assoc(S^c)\right)}$$

From theory is known that this value, should be low. Values found by the two algorithms:

Spectral clustering	***PageRank Nibble***
0.019	0.0288

Table 3. *Minimum conductance results*

Based on the minimum conductance value though, the spectral clustering approach is shown to yield a better network partition.

4. Link Prediction

For this task the *Local Path* and *Katz* methods were used.

The link prediction task, on a dataset taken from a social network, might be used to suggest pages, other pages they might want to connect with. The link prediction is not an easy task, meaning that sometimes the suggestions/predictions are accurate and helpful, while some others, are not.

Two pages are likely to connect to each other based on a similarity matrix S:

$$S_{Katz} = \sum_{l\geq 1} \beta^l A^l$$

$$S_{LP} = A^2 + \beta A^3$$

The performance is evaluated through AUC and Precision scores. The test was performed with:

- β=0.01
- Test set = 80%
- Probe set = 20%

and the following results were obtained:

	Katz	***Local Path***
AUC	0.77854	0.7725
Precision	0.153	0.146

Table 4*. Results obtained from the two approaches*