Athens University of Economics and Business Dept. of Informatics
M.Sc. Program in Business Analytics
Course: Social Network Analysis Spring 2019
Instructor: Katia Papakonstantinopoulou

Homework 2:
From raw data to temporal graph structure exploration

Due 11:59pm EEST June 23, 2019

General Instructions

Your answers should be as concise as possible.

Submission instructions: You should submit a compressed directory, containing your answers and code, via `https://e-mscba.dmst.aueb.gr`.

Submitting answers: Prepare a report with your answers on this homework in a single PDF file named *hw2.pdf*

Submitting code: Prepare the source file(s) with your code.

Problem

1 Twitter mention graph

Your first task is create a weighted directed graph with igraph,[1] using raw data from Twitter. You will download a compressed file with Tweets posted during July 2009. [2] The format of the file is the following:

```
T 2009-07-01 00:04:20
U http://twitter.com/greeneyed_panda
W @Dprinzessin jajajajajajaja..... no.....
```

In the tweet above, `T` indicates the time the tweet was posted (2009-07-01 00:04:20), `U` indicates the user that posted it (*greeneyed_panda*) and `W` is the text

[1] `https://igraph.org/r/`
[2] `https://drive.google.com/open?id=1RjWUg-6KrVOjJPZHHQg-h_9gSSWZUPn-`

of the tweet (@Dprinzessin jajajajajajaja..... no.....). Twitter subscribers can use the '@' character to make a mention to another subscriber, e.g., in the above tweet, user *greeneyed_panda* made a mention to user *Dprinzessin*.

You will first manipulate the raw data with the programming language of your choice to create a total of 5 .csv files, one for each of the first five days of July 2009, using the following format:

```
from,to,weight
user1,user2,5
user2,user1,1
user1,user3,2
...
```

Each .csv file should describe the weighted directed mention graph for the respective day, e.g., in the example above *user1* has made 5 mentions to *user2*, *user2* has made 1 mention to *user1*, and *user1* has made 2 mentions to *user3*.

Having created the .csv files it will be trivial to use them and create the respective igraph graphs.

Your submission should include the code you used to create the .csv files (any programming language), the code you used to create the igraph graphs (R) and the 5 (compressed) .csv files.

2 Average degree over time

Your next task is to create plots that visualize the 5-day evolution of different metrics for the graph. More specifically, you will create plots for:

- Number of vertices
- Number of edges
- Diameter of the graph
- Average in-degree
- Average out-degree

What do you notice for each of the 5 above metrics? Are there significant fluctuations during these five days?

3 Important nodes

Next, you will write to code to create and print data frames for the 5-day evolution of the top-10 Twitter users with regard to:

- In-degree
- Out-degree
- PageRank

Again, provide short comments on your findings. Do you notice variations on the top-10 lists for the different days?

4 Communities

Your final task is to perform community detection on the mention graphs. Try applying fast greedy clustering, infomap clustering, and louvain clustering on the undirected versions of the 5 mention graphs. Are you able to get results with all methods? Include a short comment on your report regarding the performance of the 3 algorithms.

Then, pick one of the three methods as well as a random user that appears in all 5 graphs and write code to detect the evolution of the communities this user belongs to. Do you spot similarities in the communities?

Finally, you will create a visualization of the graph using a different color for each community. Make sure to have a look at the sizes of the communities and filter out all nodes that belong to very small or very large communities, in order to create a meaningful and aesthetically pleasing visualization.