CS3907/CS6444 Big Data and Analytics

Class project #1

R and Graph Analytics

Due Date: September 24, 2018 COB

### Description

1. Data Set: (on Blackboard)

Collaboration Network of Astrophysicists’ Papers in ArXiv

**Data Set:**

**astrocollab.RData** Co-authorship network between scientists posting preprints on the Astrophysics E-Print archive.

**Description**

The collaboration network of scientists posting preprints on the astrophysics archive at <http://www.arxiv.org>, 1995-1999, as compiled by M. Newman. The network is weighted, with weights assigned as described in the original papers. Please cite, as appropriate, one or more of:

M. E. J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA 98, 404-409 (2001). M. E. J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, Pys. Rev. E 64, 016131 (2001). M. E. J. Newman, Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, Phys. Rev. E 64, 016132 (2001).

**Metadata**

Edge attribute 'weight' Edge weights, based on the number of common papers and the number of authors of these papers. See the publication(s) for the definition. Vertex attribute 'name' Author name.

………..

2. Install the igraph package from one of the CRAN mirrors. You may also use igraphdata package and rgraph (included in the SNA package) as well.

3. You will have to determine how to load the data into a data structure usable by the graph packages.

I have already given you some hints, such as read\_line and read\_table.

Note that the books in Outline:TechnicalBooks might be useful.

Here’s a procedure for creating a graph from a file. Note that your data is already loaded into R data structures. You will need to determine what type of data structures they are and modify this procedure as appropriate. ASTROCOLLAB is a matrix, so you need to extract data from it.

Creating a graph

Here is some help:

1. edges1<-read.table(<some file>)   assuming you have set working directory

2. convert to matrix:  em <-as.matrix(edges1)

3. extract vectors: v1 <- em[<$rows>:1], v2<-em[#rows:2] where you need to find the number of rows in the matrix. This gets the two vertices as separate vectors

4. relations<- data.frame(from=v1,to=v2)

5. g<-graph.data.frame(relations,directed=TRUE)  need to have installed igraph

6. plot(g)

This should work, Then you can apply various functions to it.

REMEMBER: Some functions work on matrices, some on data frames, and some on both.

You may have to do this to get astrocollab to work with igraph:

* astrocollab <= upgrade\_graph(astrocollab)

Why? Astrocollab created an earlier version of R. Internal structure may vary across versions.

4. Experiment with 10 of the functions that I have shown in the lecture notes and associated PPT file on Blackboard. Present the results in your writeup.

This is a very large data set. You may have to simplify the graph somewhat in order to execute this project. If so, describe how you simplified the graph. You may use the *simplify* function, but you may have to do more than that.

5. Explore other functions in the igraph package – at least 15 of them not shown in the lecture notes. You may have to do some programming in R. There are numerous books posted on the Blackboard.

6. Determine the (a) central person(s) in the graph, (b) longest path, (c) largest clique, (d) ego, and (e) betweenness centrality and power centrality.

a. Is there more than one person with the most degrees?

b. Are there multiple longest paths?

c. Are there multiple cliques?

d. Are there more than one person with the highest ego?

e. What is the difference in betweenness centrality vs. power centrality for the cases you find? Consider comparing the nodes that are members of each set. Are there common nodes?

In each case what do you think the data tells you?

7. Find the 20 nodes with the largest networks of distance 1. Do any of these circles overlap?

a. Build a matrix of 20 nodes versus their feature vectors

b. Determine which of the 20 nodes share common authors and, for each author, list the nodes that share that author.

8. Deliverables: You will deliver, by putting a zipfile in your group’s Blackboard file, with the following naming convention: Group-N-Project-1.zip, where N is your group number. Your deliverable should encompass the following items:

* A listing of all R functions that you have written
* Demonstrations of the igraph functions that you have explored per #, and #5.
* Answers for #5.

Be clear about what you are doing with each function. Identify any problems you had and how to solve them.

Remember to name and save your workspace! In your Group area would be a good place so all members can get to it.

Include in your Word document the results required

(use a CTRL-ALT-PrintScreen) to grab the screen

You may use Irfanview 4.41 or later, [irfanview@gmx.net](mailto:irfanview@gmx.net). Paste in the screen image, and copy the image as JPEG to drop into your Word document.

7. Project #1 Value: 10 points

a. Overall description and documented R functions: 1.5 points

b. Item #4/#5 demonstrations – 2.5 pts

c. Item #6 demonstration – 2.5 points

d. Item #7 demonstration: 3.5 points

Note: To prepare for working with this data set:

>install.packages(“igraph”)

…. Lots of stuff omitted ….

So, you can see there is a lot of additional code required to support this package that RStudio downloads for you.