

ECE 18-755: Networks in the Real World

HW2 Assigned 10/12/2017 DUE 10/26/2017 (midnight)

Problem 1 (20 pts)

Starting from two real world traces, namely *BC-Oct89Ext4.txt*¹ which is obtained from arrival packet sizes in LAN traffic, and *NoC_10x10_400_8KCache_4fDPac.txt*² which is obtained from data reply requests in on-chip traffic, perform the following statistical tests and comment on the results you get:

a) Download the SELFIS software from Canvas to estimate the Hurst exponents of the arrival packet sizes in *BC-Oct89Ext4.txt* and data reply requests in *NoC_10x10_400_8KCache_4fDPac.txt* (time series) via the R/S and variance-time methods. Discuss the sources of discrepancy between Hurst exponents obtained with these two methods.

From Canvas download and run the Matlab code for *detrended fluctuation analysis* (DFA) on the arrival packet sizes in *BC-Oct89Ext4.txt* and data reply requests in *NoC_10x10_400_8KCache_4fDPac.txt* (time series). Note that for a time series $x(t)$, the average fluctuation $F(n)$ is given by the following formula:

$$y(k) = \sum_{i=1}^k [x(i) - x_{ave}]$$

$$F(n) = \sqrt{\frac{1}{N} \sum_{k=1}^N [y(k) - y_n(k)]^2}$$

where n is the window (box) size, $y_n(k)$ (also called the time series trend) represents the least square linear fit of the vector $x(t)$ in the k -th window, and N is the length of the time series.

Plot $F(n)$ as a function of n using a log-log scale. More precisely, what can you infer now based on your DFA analysis with respect to the data and the first two methods used to quantify the Hurst exponent?

-
1. The first and the second column represent the arrival time and the size of the received packet at a LAN observation station, respectively. More details are at: <http://ita.ee.lbl.gov/html/contrib/BC.html>
 2. The first and the second column represent the arrival time and the response time for a data request (data reply request) in a 10x10 mesh NoC running a Perl scripting application. Both are measured in clock cycles.

Problem 2 (40 pts)

The goal of this problem is to explore LinkedIn ego-network structure using provided data. The data contains 250 ego-networks which have user attributes and their relationships (with types). Each student is randomly assigned an ego-network as shown on Canvas. We define the *ego-user* to be the person (*i.e.*, node) whose ego-network is analyzed. Each ego-network dataset comes with the following information:

- a. Complete edge list for the ego-network

Example:

U0 U0 U1

U0 U0 U2

...

U1046 U1046 U1172

U1046 U1046 U1173

Here, a record "U0 U0 U1" means that U0 is connected to U1 in U0's ego-network.

- b. Relationships between ego users' and their friends

Example:

U3297 U3299 L4

U10305 U10351 L2

U10305 U10369 L2

Here, the record "U10305 U10369 L2" means that the type of relationship between U10305 and U10369 is L2. Of note, there can possibly exist 8 types of relationships (total) in your ego-network.

- c. Profile attributes (location, college, employer) for the users

Example for

college.txt

U346

2

C0

C1

Here, U346 is UserID, 2 is the number of values associated with the attribute, e.g., U346 attended two colleges (C0 and C1).

For more details of the dataset, please refer to <https://arxiv.org/abs/1309.4157>

In this problem, we want you to perform an in-depth analysis of your assigned ego-network. The following steps will allow you to understand the structure of ego-network in detail.

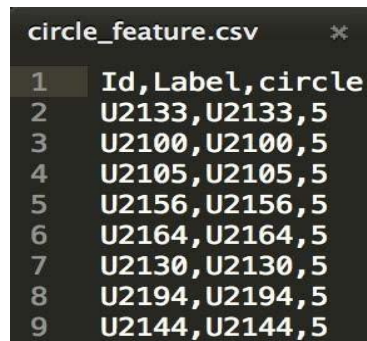
1. Preliminary analysis using Gephi

- a. **Visualization.** Visualize your data using at least two layout techniques. Choose the edge to be *directed*. Color the ego-user in your network and the edges connecting to it in *red*. Report the number of nodes and edges in your ego-network.
- b. **Network analysis.** Report degree distribution, node degree, clustering coefficient, average path length, betweenness centrality distribution, and number of communities.

2. One step further

- a. **Popularity.** Who is the most popular person in your ego-network? Explore node degree and betweenness centrality in your ego-network with the ego-user *removed*. Does the most popular person score high on both metrics? Highlight the most popular people in your network (again, the ego-user should be removed in this case).
- b. **Community structure.** In this part, we want you to visualize communities using different methods:
 - i) Use the modularity-based community detection embedded in Gephi to detect the communities in your ego-network. Visualize and color communities using different colors.
 - ii) Explore the type of relationships among users in your assigned ego-network. Group users together if their edge type (relationship) is the same. For instance, in example b. in the beginning of this problem, U10305, U10351 and U10369 should belong to the same group L2, while U3297 and U3299 should belong to another group L4. We call these groups the *ground truth circles*¹. Visualize your ground truth circles and color them.
 - iii) Compare the results from ii) with the communities detected by Gephi in i) and discuss similarities/differences. Are there people who act as bridges between different ground truth circles? If yes, re-size and visualize these nodes (users) using bigger weights.

Hint: To visualize these circles, import a spreadsheet (.csv file containing Id, Label, and circle of each node) into Gephi and color nodes by choosing the attribute “circle”. A snapshot of such a spreadsheet is shown below:



	Id,Label,circle
1	U2133,U2133,5
2	U2100,U2100,5
3	U2105,U2105,5
4	U2156,U2156,5
5	U2164,U2164,5
6	U2130,U2130,5
7	U2194,U2194,5
8	U2144,U2144,5
9	

- c. **Clustering.** Compute the average clustering coefficient for each ground truth circle in your network and compare it to the average clustering coefficient of the entire ego-network. Do you see any significant differences?

1. Note that although there may exist 8 types of relationship among users, some type of relationships may be missing in your assigned ego-network, and some edges may not be labeled as well, so just work on whatever is available for your ego-network.

d. **Homophily**¹. Explore various features (*e.g.*, college, employer, location) of users in different ground truth circles. Use entropy² to measure the homophily of each circle under different features, and then compare it with the average clustering coefficient of each circle.

Report your results in a table as shown below. Replace ‘*’ with the value of clustering coefficients or entropy measured homophily; fill in the corresponding row(s) as ‘NA’ if your ego-network does not have certain circle(s).

	<i>clustering coefficient</i>	<i>Homo college</i>	<i>Homo employer</i>	<i>Homo location</i>
<i>circle 1</i>	*	*	*	*
<i>circle 2</i>	*	*	*	*
<i>circle 3</i>	NA	NA	NA	NA
<i>circle 4</i>	*	*	*	*
<i>circle 5</i>	*	*	*	*
<i>circle 6</i>	*	*	*	*
<i>circle 7</i>	*	*	*	*
<i>circle 8</i>	*	*	*	*

3. Getting even deeper

- How homogeneous are communities in your ego-network? Do they tend to share similar characteristics? Do similar people belong to the same communities?
- How interconnected is your LinkedIn ego-network? Is there a high degree of overlap between communities in your ego-network?

Problem 3 (10 pts)

Consider the adjacency matrix of a protein interaction network stored in *pin_hsapiens.dat* which is available on Canvas. Download the MEMB code and compute the fractal dimension³ of the network. What conclusion can you draw from your analysis?

-
- Homophily (*i.e.*, "love of the same") is the tendency of individuals to associate and bond with similar others.
 - Entropy is defined as $-\sum_i p(x_i) \log p(x_i)$, where $p(x_i)$ is the probability of feature x_i belonging to users in that circle.
 - To compute the fractal dimension of a complex network you have to compute iteratively the number of boxes N_b (with a certain box size l_b) required to cover the entire graph. For more details read the *Fractal_dimension.pdf*.

Problem 4 (15 pts)

In this problem, you need to analyze the community structure in research collaboration. Towards this end, you need to follow a few steps: First, you need to download the *arXiv* collaboration network used in Homework 1 (i.e., *arXiv_lcc.gml*). Second, you need to use the community detection algorithm provided in Gephi (<https://gephi.org/>) to identify the communities in this network. You need to supply a wide range of resolutions (a user-specified parameter) to obtain the resulting modularity and number of communities. Note that, in contrast to the case of *granularity* mentioned in class, a lower *resolution* in Gephi's community detection algorithm corresponds to a higher number of communities.

- a) Plot the modularity as a function of number of communities.
- b) Using modularity resolutions of 0.1 and 0.75, visualize this network while coloring each node according to the community it belongs to. Based on your visualizations, does the result of the community detection algorithm make intuitive sense? Why or why not?

Problem 5 (15 pts)

In this problem we will analyze the connectivity of transportation systems, more specifically the Light Rail and bus routes of Pittsburgh. You will find CSV files on Canvas that have the station information for both lines of the T Light Rail, and the stops for the 28X bus line. The entries are listed in the sequential order of stops; provided are maps for reference.

- a) Visualize the transportation network, referencing the 'Station Name' field for node identification.
- b) Analyze the 'small-worldness' of the network you obtain by calculating the average path length, nodes degree, and clustering coefficient.
- c) Insert several long-distance links (do not connect neighboring nodes) to create 'triangles' within the network in order to reduce the path length by at least 30%. Comment on your results and explain them within the context of 'small-world' networks.

Note on submission:

For all problems in this homework, everything is handled electronically. Prepare your answers using either Word or Latex and **create a single PDF file for submission** (i.e. this PDF file should contain the answers to all the problems in the HW). Also, put the source code and all related files for each problem in a *separate folder*. Finally, compress everything and name it as "yourAndrewId_hw2.zip", and deposit it on the Canvas under Assignments > Homeworks Fall 2017 > HW2 Submission.

* Work individually on all problems.

Good luck!