TOBB ETÜ, BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

# BİL 403/503 Social Networks

## HOMEWORK 1 (60 Points)

## Due Date: February 18, 2021

In this homework, you will be implementing some of the algorithms we have seen in class on a public dataset in order to experimentally verify some of the notions related to the structure of social networks. **Your programs are to be written in Java.** In your experiments, you will be using a collaboration network on the theory area of the High Energy Physics based on the papers submitted to arXiv. The dataset you will use is available at SNAP High Energy Physics - Theory collaboration network.

**When solving the assignment you can NOT consult or get help from any person, web page, book etc. other than the textbook.**

## 1 [20 POINTS] THE SMALL-WORLD PHENOMENON

Your first experiment is to check whether the small-world phenomenon holds on your dataset. For that, you should first find the giant component in the dataset. Then, you should compute the distance between every pair of vertices in the giant component. Notice that if there are $n$ vertices in the giant component then there are $\frac{n\times(n-1)}{2}$ such pairs. You should produce a histogram of the distances of the pairs of vertices in the giant component. So, in your plot the x-axis should be distances (small integers $1,2,\ldots,18$) and the y-axis should be the number of pairs of vertices in the giant component (with a distance given in the x-axis).

## 2 [40 POINTS] GRAPH PARTITIONING

In this question, you will implement the Girvan-Newman method of graph partitioning on the giant component of your dataset until the giant component is no longer a single connected components. Notice that at each iteration the Girvan-Newman method is removing

all the edges with the same betweenness value so it is possible that the giant component be divided into more than two connected components in the same iteration. Then, you will sort the vertices in the same (newly found) connected components by node id.

**Submission Details:** Create a folder whose name is your student number. Then create 3 folders named: Q1, Q2, and, dataset. Put the dataset that you downloaded (extracted) inside the dataset folder and your source codes inside the respective folders. Only put the .java files into Q1, Q2. (Do not put class or txt files). If you use the same data structures (classes) in both question please copy them into both of folders. Then zip the main folder and send a single .zip file to uuacikalin+bil403hw1@gmail.com.

Your programs should produce a text file named output1.txt and output2.txt for Q1 and Q2, respectively. For the first question, the output1.txt should contain the histogram. For the second question, the output2.txt should contain the connected components each in separate line. The class with the main method should be named Main. Your programs should be executable by the following lines:

```
unzip YourStudentNumber.zip
cd YourStudenNumber/Q1
javac *.java
java Main
cd ../Q2
javac *.java
java Main
```

**Furthermore, for the second question, your program should not take more than** 30 **minutes to finish in my computer, otherwise, you will receive** 0 **points. If use an efficient implementation it should take around** 10 **minutes. So, it should not run for hours in your computer. The student with the fastest implementations for the second question will be rewarded** 10 **additional points.**