Social Media Analysis (SS2016) Dr. Krishna P. Gummadi

Exercise Sheet 6 – Community Detection II

Max Planck Institute for Software Systems / Saarland University

Instructions

1. You have to submit this assignment (either digitally or on paper) by 14:00 on 12th July 2016.
2. Irrespective of your submission preferences (digital or paper), you have to email the code of the coding assignments to the instructors.
3. To submit your assignments digitally, email your solutions to sma-ss16-instructors@mpi-sws.org, with the subject as "[SMA] HW 6 : <last name 1> <matriculation number 1>, <last name 2> <matriculation number 2> solutions".
4. Add your full name(s) and the matriculation number(s) in the beginning of all types of assignment solution submissions (in digital copy, on paper, inside the code as comment etc).
5. **You can submit solutions to both section A and section B. Your grade for this exercise will be maximum of the grade for section A and grade for section B)**. For example if you score 5 in section A and 0 in section B, your final grade for this exercise will be 5.
6. Instructions for coding assignment:
   (a) You are free to use any programming language you are comfortable with. A reference for an efficient way to implement graph classes and methods is the github code base for snap at `https://github.com/snap-stanford/snap`.
   (b) You should be able to run your code in a computer with a reasonable configuration (for instance 2GB or more RAM).
   (c) Please include the results outputted from your code into your solution sheet as described in the respective tasks. Additionally, you should also submit your code and instructions on how to run your code in a text file (howtorun.txt) via email to the instructors. If there are any additional resources (eg. library) that are required to run your code, please mention those in the how to run file.
   (d) We should be able to run the submitted code on our machine. If it fails to run on our machine, you will not get any marks for the coding part of the assignment.
   (e) If any part of your code takes a long time to run (e.g., more than 10 minutes) report that in the instruction file with an estimate of time required.

Section A

Dataset Description

For this assignment you are provided a *new* graph of the Facebook network of 1,219 Rice University students, with the following sets of information:

A. **Social-links of students**: The friendship graph of the students is undirected and unweighted. Each line in the .elist file contains two anonymized student IDs, which indicates that a friendship link exists between the two students. It can be downloaded from `http://courses.mpi-sws.org/sma-ss16/assignmentData/community-detection/rice-univ-facebook-links_new.elist.txt.gz`. (Note that this is different from the graph given in Assignment 5.)

B. **Attributes of students**: Additionally, for each student, three attributes are specified in the .attr file and it can be downloaded from `http://courses.mpi-sws.org/sma-ss16/assignmentData/community-detection/rice-facebook-undergrads-users.attr.tar.gz`. The three attributes are:

   i. *College*: The second column in the .attr file denotes the college ID for the student, which is in the range 1-9.

   ii. *Age*: The third column in the .attr file denotes the age of the student, which is in the range 18-22.

   iii. *Major*: The fourth column in the .attr file denotes the major course ID for the student, which is in the range 1-60.

   You will need to use these data files for solving the problems in this exercise sheet.

Problem 1 (20 marks)

In this exercise, we will examine the *semantic groups* identified based on the different attributes. For each attribute, say college, we can define a partition of the graph, such that all the students who share the same college ID are in the same (semantic) group. In this case we would have 9 semantic groups based on the attribute college and these 9 groups together would constitute a *partition* of the graph based on the attribute college. Similar partitions based on semantic groups can also be constructed for the other two attributes of age and major.

Your task is to evaluate how well connected are the students within these semantically identified partitions and groups.

Task 1: Global partitions & their quality (10 marks)

The measure of the quality of a global partition – **modularity** – is defined as:

$$Modularity = \frac{1}{2m}\left(\sum_{i,j \in V} (A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j)\right) \tag{1}$$

$$\delta(C_i, C_j) = \begin{cases} 1, & \text{if } C_i{=}C_j. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where A is a adjacency matrix of the graph, C_i is community to which node i belongs, k_i is the degree of node i, m is the total number of edges, and V is the set of nodes.

Using this measure of modularity, we will evaluate the quality of partitions based on the three attributes.

Write code to compute the modularity values for the three partitions that are identified by the attributes: college, age, and major. Your code should take as input the .elist and .attr files for the Rice University network, and output the modularity values for the partitions based on each of the attributes.

1. Submit your code file named as **semantic-groups-modularity.py** (or .cpp or any other programming language you prefer). You should also submit an instructions file named **semantic-groups-modularity.howtorun.txt** to tell us how to compile and run your code. The output of your code should be the modularity values for the three partitions.
2. Use the output of your code to fill and complete Table 1, and add the table to a pdf file named **semantic-groups.pdf**.
3. Based on the values in the completed Table 1, briefly comment on which of these attributes results in the best partitioning of the Rice University graph and whether you have expected it.

Partition based on	**Modularity**
College	
Age	
Major	

Table 1: Modularity values computed for the partitions based on the three attributes : college, age, and major

Task 2: Local communities & their quality (10 marks)

The measure of the quality of local communities – **conductance** – is defined as:

$$Conductance = \frac{e_{A,G-A}}{\min(e_A, e_{G-A})} \tag{3}$$

where G is the full graph while A is the community that we are computing the conductance value for. Then $e_{A,G-A}$ is the number of edges between A and $G - A$, and e_A is the number of edges between A and G.

Using this measure of conductance, we will evaluate the quality of every semantic group that is identified using the three attributes. These semantic groups are shown in the first column on Table 2.

Write code to compute the conductance values for the semantic groups that are identified for the different values of the three attributes: college, age, and major. Your code should take as input the .elist and .attr files for the Rice University network, and output the conductance values for the semantic groups based on each of the three attributes.

1. Submit your code file named as **semantic-groups-conductance.py** (or .cpp or any other programming language you prefer). You should also submit an instructions file named **semantic-groups-conductance.howtorun.txt** to tell us how to compile and run your code. The output of your code should be the conductance values for the semantic groups based on the attributes : *college (9 groups), age (5 groups), and major (52 groups).*

Semantic Group	Conductance
College 1 College 2 ... College 9	
Age 18 Age 19 ... Age 22	
Major 1 Major 2 ... Major 60	

Table 2: Conductance values for the different semantic groups identified based on the three attributes : college, age, and major

2. Use the output of your code to fill and complete Table 2, and add the completed table to the aforementioned pdf file named **semantic-groups.pdf**.

3. Based on the values in the completed Table 2, summarize in 3-4 sentences any interesting observations that you can make about the connectedness of the students in the different semantic groups.

Problem 2 (10 marks)

Background & Motivation: Many applications for social media leverage the attributes of the social media users to give meaningful services. A typical example of such applications are the information retrieval systems like search and recommendation systems. For example, to pick a post for a user to read, the recommendation can be based on the attributes of the user like her age, gender or profession.

But many times on social networks, different people provide information about different subsets of their attributes. For example, on Facebook while some members provide complete profile information such as where they study or live, how old they are, and where they work, there are many other members who dont give out the complete information. Therefore, mechanisms need to be developed to predict or guess the missing attributes for these users. One of the methods could be to figure out which community do these users fall in and to use the common attributes of the members of this community to guess the missing attributes of these target users.

Problem Statement: In this exercise you are given the social links of 5 additional students (on top of the Rice University students' social graph and attributes described at the beginning of the tasksheet), but their attribute information is not unknown. Your task is to predict the most likely values of the three missing attributes for these 5 additional students.

The main intuition behind solving the problem is that the semantic groups of students that you identified in Problem 1 can also be used to predict the missing attributes information of these new students. For example, in order to predict the most likely age of a new student, you could evaluate with which age based semantic group is this new student best connected, such that its addition to the group leads to a *"better"* conductance value for the group (as compared to the conductance values computed in Problem 1 - Task 2).

Data: The student IDs of these 5 students are 608, 1583, 2449, 3915 and 4186. Their social links file can be downloaded from `http://courses.mpi-sws.org/sma-ss16/assignmentData/community-detection/prediction-task-users-links.elist.tar.gz`, and the file format is "StudentID: friend1, friend2,...".

Write code to *predict the top 2 most likely values for the missing attributes* (college, age, major) for these 5 students.

1. Submit your code file named as **predicting-missing-attributes.py** (or .cpp or any other programming language you prefer). You should also submit an instructions file named **predicting-missing-attributes.howtorun.txt** to tell us how to compile and run your code. Your code should take as input the Rice University social graph, known attributes of the students and the social links of the 5 new students. For every additional student, your code should output the new conductance value for each semantic group (from Problem 1) when this additional new student is added to the group.

2. Use the output of your code to fill and complete five instances of Table 3 (one for each new student) and add the tables in a pdf file named **predicting-missing-attributes.pdf**. Here "Original Conductance" is the value for conductance for a semantic group G (computed in Problem 1) and "New Conductance" refers to the new value of conductance when an additional student A is added to the group G.

   In the title of each table mention the new student's ID for whom the table is constructed.

Semantic Group	**Original Conductance**	**New Conductance**
College 1 College 2 ... College 9		
Age 18 Age 19 ... Age 22		
Major 1 Major 2 ... Major 60		

Table 3: For an additional student A, the conductance values for the original semantic group G and the new group comprising of G and the new student A

3. At the end of the table for each of the 5 new students, state your *top 2 most likely predicted values* for the three attributes: college, age and major. For instance, for the attribute age your top 2 most likely predictions could be 18 (most likely) and 19 (second most likely) based on the conductance values in the table.

   Also briefly summarize your reasoning for choosing these to be the two most likely predictions, using the conductance values in the table.

   NOTE: We will reveal the actual attribute values of these 5 students after this assignment, so that you can evaluate how good were your predictions.

Section B

Problem 3 (Research, 30 marks)

Identify a novel and significant problem based on the topics discussed in the class so far. The problem may be about devising some new model, proving a new theorem, identifying a new application or any other idea that you might have. You must first clearly identify the problem you are interested in solving, then motivate why you think this problem is important to answer and non-trivial to solve. Finally, you must propose / sketch a solution that you have in mind to address this problem.

Note that this part of the Exercise is aimed for the students who enjoy research and are motivated to identify and solve problems that remain unaddressed in today's world.

The grading for this section will be highly subjective and will be judged based on the novelty of the identified problem, motivation of the problem and the solution sketch that you provide. If you identify a problem that also interests us, we would be happy to follow it up with you. Additionally, we would invite some students who propose an exceptional problem, to present their work briefly in one of the tutorials.