Big Data Processing: Assignment 3
Deadline: 16.04.24, 11.55 pm IST
Marks: 20

You are given an undirected unweighted graph. Your goal is to write **pyspark program** for the following two problems.

1. Count the number of heavy hitter nodes in the graph.
2. Count the exact number of triangles in the graph.

Submit a single pyspark program (.py) which will print on screen the above outputs in the following format.

No of heavy hitter nodes: n1
No of triangles: n2

Path to data file: The data file is attached in moodle.
Data format: The graph file is a two-column file where each line of the file denotes an edge.

We will evaluate your program on a linux system from command line with the arguments as follows:

spark-submit <your-code> <path to file>

Submission guidelines:

You need to submit the program as a single python file in moodle. The file name must follow the format: **assignment-3-roll.py** (where the roll denotes your roll number in capital letters that must match exactly with your IITKGP roll number). Please note that if you fail to follow the format, your program may not be evaluated at all.

Important notes:

1. No credit will be given if your program does not run and produces wrong output.
2. No credit will be given if your program does not implement map with sort and shuffle and does not adhere to the specified buffer size.
3. No submission will be accepted after deadline.
4. It is your responsibility to check that the file has been submitted successfully.
5. Plagiarism from friend or from web will invite negative (**-10)** marks.