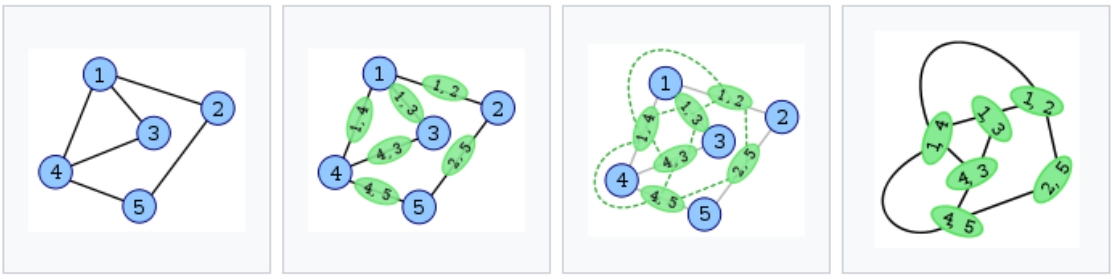
COMP9313 2018s2 Assignment

Question 1. MapReduce (5 pts)

Problem Background: Given an undirected graph G, its "line graph" is another graph L(G) that represents the adjacencies between edges of G, such that:

- each vertex of L(G) represents an edge of G; and
- two vertices of L(G) are adjacent if and only if their corresponding edges share a common endpoint ("are incident") in G.

The following figures show a graph (left) and its line graph (right). Each vertex of the line graph is shown labelled with the pair of endpoints of the corresponding edge in the original graph. For instance, the vertex on the right labelled (1,3) corresponds to the edge on the left between the vertices 1 and 3. Vertex (1,3) is adjacent to three other vertices: (1,2) and (1,4) (corresponding to edges sharing the endpoint 1 in G) and (3,4) (corresponding to an edge sharing the endpoint 3 in G).



Graph G | Vertices in L(G) constructed from edges in G | Added edges in L(G) | The line graph L(G)

Problem: Given you the adjacency list of an undirected graph G, use MapReduce to generate the adjacency list of its line graph L(G). Note that each edge connecting two nodes i and j is represented by (i, j) in L(G) (if i<j). In the output, the edges in each list should be ranked in ascending order by comparing the first node and then the second node. Write the pseudocode for this problem, and consider the efficiency of your solution.

Take the above figure as an example, sample input and output are as below:

Input:	Output:
1: 2, 3, 4 2: 1, 5 3: 1, 4 4: 1, 3, 5 5: 2, 4	(1, 2): (1, 3), (1, 4), (2, 5) (1, 3): (1, 2), (1, 4), (3, 4) (1, 4): (1, 2), (1, 3), (3, 4), (4, 5) (2, 5): (1, 2), (4, 5) (3, 4): (1, 3), (1, 4), (4, 5) (4, 5): (1, 4), (2, 5), (3, 4)

Question 2. LSH (5 pts)

(i) Given two documents A (“the sky is blue the sun is bright”) and B (“the sun in the sky is bright”), using the ***words*** as tokens, compute the 2-shingles for A and B, and then compute their Jaccard similarity based on their 2-shingles.

(ii) We want to compute min-hash signature for the two documents A and B given in Question (i), using two pseudo-random permutations of columns using the following function:

$h1(n) = 5n - 1 \bmod M$

$h2(n) = 2n + 1 \bmod M$

Here, n is the row number in original ordering of the 2-shingles (according to their occurrence in A then B), and M is the number of all 2-shingles you have computed from A and B. Instead of explicitly reordering the columns for each hash function, we use the implementation discussed in class, in which we read each data in a column once in a sequential order and update the min hash signatures as we pass through them.

Complete the steps of the algorithm and give the resulting signatures for A and B.

Submission:

Deadline: Sunday 4th November 11:59:59 PM

Please provide your solutions to these questions in a pdf file named as "answers.pdf". Log in any CSE server (williams or wagner), and use the give command below to submit your solutions:

$ give cs9313 assignment answers.pdf

Or you can submit through:
https://cgi.cse.unsw.edu.au/~give/Student/give.php

If you submit your assignment more than once, the last submission will replace the previous one. To prove successful submission, please take a screenshot as assignment submission instructions show and keep it by yourself.

Late submission penalty

You will receive zero marks for this assignment.

Plagiarism:

The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such an offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Relevant scholarship authorities will be informed if students holding scholarships are involved in an incident of plagiarism or other misconduct.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this subject. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted you may be penalized, even if the work was submitted without your knowledge or consent.