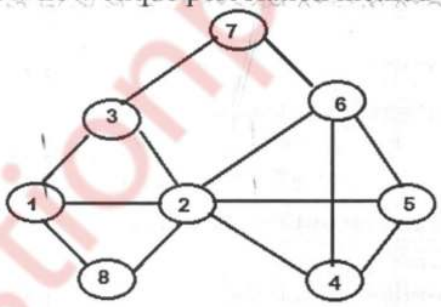
BDA numericals

Big Data Analytics (University of Mumbai)

Q.3. (a) Suppose a data stream consists of the integers 1,3,2,1,2,3,4,3,1,2,3,1. Let the Hash function being used is h(x) = (6x+1) mod 5; estimate the number of distinct in this stream using Flajolet - Martin algorithm. **(10)**

(b) For the given graph show how clique percolation method will find cliques. **(10)**

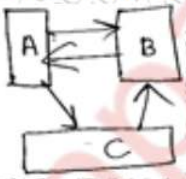


Q.5. (a) Consider the web graph given below with six pages (A, B, C, D, E, F) with directed links as follows. **(10)**

A→ B,C
B→A,D,E,F
C→AF

(b) Find the jaccard distance and cosine distance between the following pairs of set: X=(0,1,2,4,5,3) and Y=(5,6,7,9,10,8).

3. a) Compute simplified page rank using damping factor d = 0.9 for web. **(10)**



5. a) Define Hub and Authority. Compute Hub and Authority scores for web. **(10)**

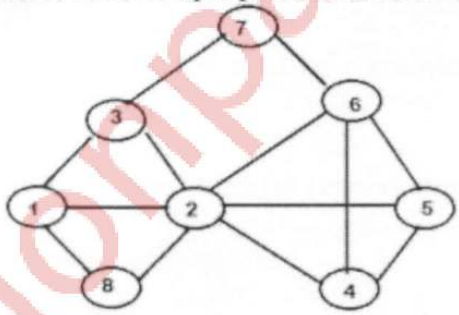
D) Find Cosine Distance between the d1 and d2 vectors:

Index	1	2	3	4	5	6	7	8	9	10
d1	5	2	1	0	0	0	0	1	3	7
d2	5	2	1	0	0	1	2	2	0	2

(b) Find the jaccard distance and cosine distance between the following pairs of set: X=(0,1,2,4,5,3) and Y=(5,6,7,9,10,8). (5)

Q.3. (a) Suppose a data stream consists of the integers 1,3,2,1,2,3,4,3,1,2,3,1. Let the Hash function being used is h(x) = (6x+1) mod 5; estimate the number of distinct in this stream using Flajolet - Martin algorithm. (10)

(b) For the given graph show how clique percolation method will find cliques. (10)



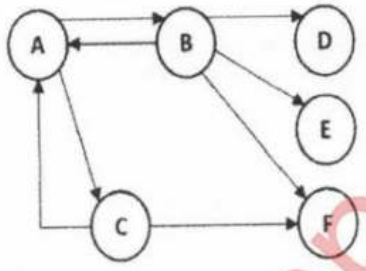
Q.5. (a) Consider the web graph given below with six pages (A, B, C, D, E, F) with directed links as follows. (10)

A→ B,C
B→A,D,E,F
C→AF

Assume that the PageRank values for any page m at iteration 0 is PR(m)=1 and teleportation factor for iterations is β=0.85. Perform the page rank algorithm and determine the rank for every page at iteration 2.

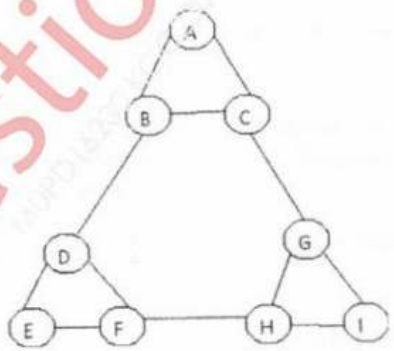
b) i) Find Jaccard distance {1, 2, 3, 4} &{2, 3, 5, 7} and {a, a, a, b}& {a, a, b, b, c} **10**
ii) Find Hamming Distance between 110011 & 010101 and 11001 & 01011
iii) Compute the cosines of the angles between (3,−1, 2) and (−2, 3, 1).

(b) Define PageRank. Using the web graph shown below compute the PageRank at every node at the end of the second iteration. Use teleport factor = 0.8. **(10)**



Q. 6 **(a)** Explain clearly with diagrams how the PCY algorithm helps to perform frequent itemset mining for large datasets. **(10)**

(b) For the graph given below use betweenness factor and find all communities **(10)**



Q. 4 **(a)** Suppose a data stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Let the hash function being used is h(x) = 3x + 1 mod 5; Show how the Flajolet-Martin Algorithm will estimate the number of distinct element in this stream. **(10)**

Qu-2 a. Write a Map-Reduce Algorithm for Binary search tree. Explain the flow of execution. [10]

Qu-2 b. Suppose a stream consists of the integers 2,1,6,1,5,9,2,3,5. Let the hash functions all be of the form h(x)=ax+b mod 16 for some a & b. You should treat the result as a 4 bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is : [10]

a) $h(x) = 2x + 3 \bmod 16$

b) $h(x) = 4x + 1 \bmod 16$

c) $5x \bmod 16$

Qu-3 a. Explain Different types of recommendation system with real time examples. [10]

Qu-3 b. Consider the portion of a Web graph as shown in Figure-1 [10]

i) Compute the hub and authorities scores for all nodes

ii) Does this graph contain spider traps? Dead ends? If so, which nodes

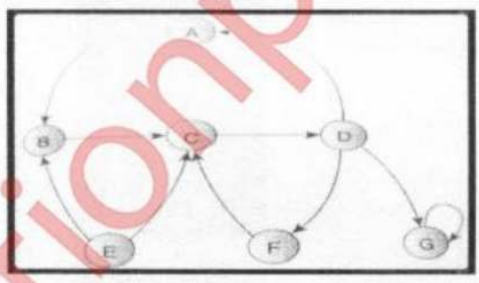


Figure-1 Web graph