1. [1 point] Which of the following statement(s) is(are) INCORRECT for Combiners?

A. Combiners save network time by aggregating values inside the reducers after the map process
B. Typically 1 Combiner combines the values of all keys from all the mappers (all nodes).
C. The Combiners instances are typically run on every node that runs the map tasks.
D. The Combiner can be considered a "mini-reduce" process.

2. [1 point] **Select ALL** of the problems/data suited for Map-Reduce.

A. Problems that requires faster response time, such as online purchases
B. Data set is truly big
C. Data that can be expressed as key-value pairs without losing context, dependencies
D. Problems that need some machine learning algorithms containing gradient-based learning.

3. [1 point] For the WORD COUNT algorithm that we saw in class, assume the data set of total size D (i.e. the total number of words in the data set), and number of distinct words is W, and assume there is NO combiner. **Circle** the total communication between the mappers and the reducers (i.e. the total number of key-value pairs that are sent to the reducers).

A. W
B. D
C. W + D
D. Not enough information

4. [1 point] Apply the ONE-PHASE Map Reduce algorithm to the matrix A and matrix B:

$$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 3 \\ 0 & 1 \\ 5 & 2 \end{bmatrix}$$

A B

Identify in the list below, Select ONE of the key-value pairs that is output of Map Task.

A. ((1, 1), ('A', 0, 0, 1))

B. ((1, 2), ('A', 2, 1, 4))

C. ((2, 1), ('B', 2, 1, 2))

D. ((2, 2), ('B', 1, 2, 3))

5. [1 point] How does a Map worker deal with failures?

A. Reset completed tasks to idle
B. Reset in-progress tasks to idle

C. Reset both completed and in-progress tasks to idle
D. Abort task and notify clients

6. [1 point] [True/False] In Spark, actions always return data to the driver, but transformations are executed lazily.

7. [2 points] Recall Bonferroni's principle when we looked at an example for detecting suspicious activity. Suppose that we make the following assumptions. We track 1 million people for 1000 days. Each person stays in a hotel 1% of the time. Each hotel holds 100 people and there are 100 hotels.

a. [2 points] What is the expected number of "suspicious" pairs of people (i.e., they went to the same hotel on some two days)?
   A. 250
   B. 2500
   C. 25000
   D. 250000

8. [1 point] How does Map-Reduce address the challenges that are seen in the cluster computing network?

A. Stores data redundantly on multiple nodes
B. Move data closer to computation to maximize data movement
C. Distributed programming capabilities
D. Move computation closer to data to minimize data movement

9. [1 point] Each time an action is run on an RDD, the RDD is recomputed by default.
A. True
B. False

1.[6 points] For the **A Priori algorithm**, consider the following input file of basket data, where each basket lists (i.e., { }) the items it contains. For a support threshold **s = 3**, answer the following questions.

***Basket data: {a, b, c, d, e} {d, e, c} {a, b, c, f} {a, b, c, d}**

1.1) [1 pts] In pass 1, which of these items are frequent?

Item	Count	Frequent

1.2) [2 pts] For pass 2, **which are the candidate pairs** for each basket? (Only include the pairs that **will be counted**.) (0.5 * 4 - 0.5 points if all candidate pairs in a basket are correctly identified)

For pass 2, how many candidate pairs are there for each basket? (Only include the pairs that will be counted)

Basket	Candidate pairs	# Candidate pairs
1		
2		
3		
4		

1.3) [1 pts] What is the **count for each candidate pair and** which of the **candidate pairs are frequent**?

Candidate pair	Count	Frequent

2. [2 points] Consider using the **Toivinen's** algorithm to find frequent itemsets in five items A, B, C, D and E. After the first pass we have found the following itemsets to be frequent in the sample: {A}, {B}, {C}, {D}, {B, C}, {C, D}.

2.1 [0.5 point] Please give an example of a singleton in the negative border.

2.2 [1 point] Please identify pairs in the negative border.

2.3 [0.5 point] If we found that {B, C, D} and {A, E} are not frequent in the second pass, is it safe to decide we have found all the frequent datasets?

3. [2 pts] **Use of Main Memory for Itemset Counting.** Consider the set of items: {A, B, C, D, E, ..., Z} (total **26** items). Assume integers are 4 bytes and only 1/4 of the pairs (doublets) have an occurrence > 0.

(a) [1 pt] How much space does the **triangular-matrix** method take to store the pair counts?

(b) [1 pt] How much space does the **triples** method take to store the pair counts?

Q1. Answer the following questions:[2 points]

A. How many 2-shingles does the word **"BICKERING"** have?
B. How many 2-shingles does the word **"SNICKREING"** have?
C. What is Jaccard distance between the two words?
D. What is the Jaccard similarity between the two words?
   NOTE- Write your answer as fraction ex: 2/3

Q2. Consider the following characteristic matrix of two sets: **S1** and **S2**. Suppose we use the two hash functions: **h1(x) = (3x + 2) % 9** and **h2(x) = (17x) % 7** to generate signatures of the sets as shown in the table below. [4 points]

A. Fill in the blanks for Column A and Column B.

Row	S_1	S_2	A = (3x+2) %9	B = (17x) %7
0	0	0	A0	B0
1	0	1	A1	B1
2	1	0	A2	B2
3	1	0	A3	B3
4	0	1	A4	B4
5	1	1	A5	B5
6	0	0	A6	B6
7	0	0	A7	B7
8	0	1	A8	B8
9	1	1	A9	B9

NOTE- Write your answer as A0,A1,A2…..A9 & B0,B1,B2……B9 for each of the columns.

B. Construct a signature for S_1 and S_2 based on the minhash values obtained from the h1(x) and h2(x). Fill in the blanks for the following values.

	S1	S2
h1	A	B
h2	C	D

C. What is the estimated Jaccard Similarity between S1 and S2.

D. What is the actual Jaccard Similarity between S1 and S2.(Simplify the fraction if possible)

Q3.Using the Jaccard similarities calculated from Question having the 2 hash functions h1(x) and h2(X) determine the following: is the estimate close to the actual Jaccard similarity? If not, how can the estimate be improved? [1 point]

A. Yes. There is no improvement required
B. No. The estimate can be improved by using additional hash functions to construct the signature.
C. No. The estimate can be improved by reducing the number of hash functions to construct the signature.

Q4. What is the effect of following on the false positive and false negative rate in LSH? Increasing bands *(b)*, keeping rows *(r)* constant[1 point]

a. Increase false negatives and increases false positives
b. Decreases false negatives and Increases false positives
c. Increases false negatives and decreases false positives
d. Decreases false negatives and decreases false positives

Q5. What is the correct order of finding similar items?[1 point]

a. DOCUMENT -> SET -> SIGNATURE -> SIMILARITY
b. DOCUMENT -> SIGNATURE -> SET -> SIMILARITY

c. DOCUMENT -> SIGNATURE -> SIMILARITY -> SET
d. DOCUMENT -> SET -> SIMILARITY ->SIGNATURE

Q6. In Analysis of Banding Technique probability of being a candidate pair is?[1 point]

A. A.1 - (1 - t^r)^b
B. B.(1 - t^r)^b
C. C.(1 - t^b)^r
D. D.1 - (1 - t^b)^r

Q1.
There are 2^16 documents in a repository. The word “data” appears in 2^8 documents. In a document named “Quiz”, the frequency of the word “data” is 10 and the maximum occurrence of any term in the same document is 40.
Calculate the TF.IDF score of the word “data” in the document “Quiz”. [1 pt] - 2

Q2. Select **ALL** of the statements that are TRUE about **Content-based Approach**. [1 pt]
A. It is able to recommend new & unpopular items
A. It faces cold-start or sparsity problems
A. It recommends items outside user’s content profile
A. It is unable to exploit quality judgments of other users

Q3.

	HP 1	HP 2	HP 3	TW	SW 1	SW 2	SW 4
A	5			1	4		
C				5	2	4	

A utility matrix representing ratings of movies on an 1-5 scale

Q3.1 Calculate the Cosine Similarity between user A and user C for the Features of movie rating (round to 3 decimal places) [1 pt].
Q3.2 Calculate the **normalized ratings** for user A [1 pt] and C [1 pt].

	HP 1	HP 2	HP 3	TW	SW 1	SW 2	SW 4
A							
C							

Q4. **User-based Collaborative Filtering.**

$$w_{u,v} = \frac{\sum_{i \in I}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i \in I}(r_{v,i} - \bar{r}_v)^2}} \quad P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} | w_{a,u} |}$$

	I_1	I_2	I_3	I_4
U_1	4	?	5	5
U_2	4	2	1	
U_3	3		2	4
U_4	4	4		
U_5	2	1	3	5

	I1	I2	I3	I4
U1	4		5	5
U2	4	2	1	
U3	3	?	2	4
U4	4	4		
U5	2	1	3	5

Calculate the predicted rating for U3 on item I2 (using the Pearson correlation).

Q4.1. Select all users that will be considered for calculation: [1 pt]
U1
U2
U4
U5

Q4.2 Calculate Weights:
W3,2 =? [0.5 pt]
W3,4 =? [0.5 pt]
W3,5 =? (rounded off to nearest 2 digits) [1 pt]

Q4.3 Predicted rating [1 pt]
P3,2 = ?(rounded off to nearest 2 digits)

Q5. Pearson Correlation works better than Jaccard Similarity for Item Based Collaborative Filtering (True or False) [1 pt]

Q1. Select **ALL** that are benefits of **Memory-based Approaches**. [1 pt]

A. **No feature selection is needed**
B. Can recommend an item that has not been previously rated
C. The user/ratings matrix is sparse
D. Can recommend items to someone with unique taste

Q2. **Item-based CF** leads to online systems being slower than user-based methods due to the computational complexity of search for similar items. [1 pt]:

Q3. **Collaborative Filtering** uses Product Features and User's ratings to provide recommendations such as whether a user likes/dislikes a product. [1 pt]:

Q4. Given the following description, select the corresponding **Hybrid Recommender Type**
Recommenders are given strict priority, with the lower priority ones breaking ties in the scoring of the higher ones. [1 pt]

A. Switching
B. Cascade
C. Feature Augmentation
D. Meta-Level

Q5. Extending memory-based algorithms with inverse user frequency is based on the idea that highly popular items contribute less information in similarity measures than less popular items: [1pt]:

Q6. Using **Item-based CF** (**N=3**) and the **Pearson Correlation**, calculate the rating prediction of I4 for U4 using average ratings based on **only co-rated items**.

$$w_{i,j} = \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U} (r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in U} (r_{u,j} - \bar{r}_j)^2}} \qquad P_{u,i} = \frac{\sum_{n \in N} r_{u,n} w_{i,n}}{\sum_{n \in N} |w_{i,n}|}$$

	I1	**I2**	**I3**	**I4**
U1	2	1		3
U2	3		5	2
U3		4	2	3
U4	5	3	1	**?**

Q6— Part 1. Calculate weights **using average ratings of only co-rated items**. (rounded off to nearest 2 digits):

$$W1,4 = \frac{(2-\frac{5}{2})(3-\frac{5}{2})+(3-\frac{5}{2})(2-\frac{5}{2})}{\sqrt{(2-\frac{5}{2})^2+(3-\frac{5}{2})^2}*\sqrt{(3-\frac{5}{2})^2+(2-\frac{5}{2})^2}} = \textbf{?}\ [1\text{ pt}]$$

$$W2,4 = \frac{(1-\frac{5}{2})(3-3)+(5-\frac{5}{2})(3-3)}{\dots} = \textbf{?}\ [1\text{ pt}]$$

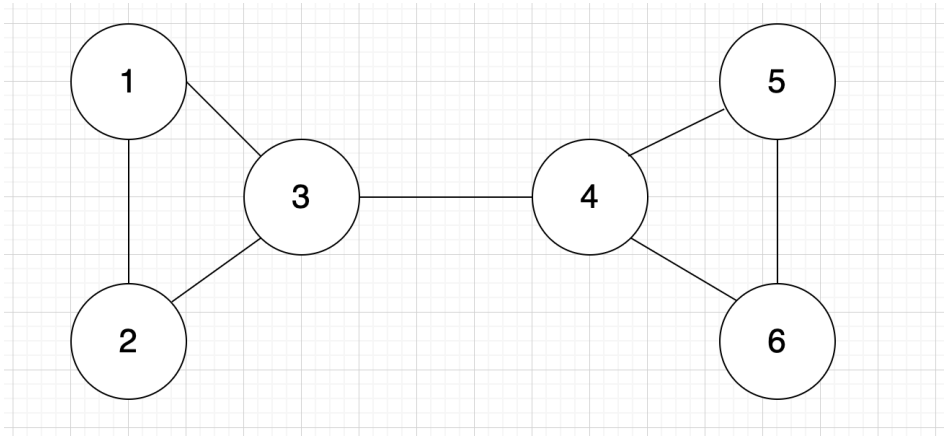
$$W3,4 = \frac{(5-\frac{7}{2})(2-\frac{5}{2})+(2-\frac{7}{2})(3-\frac{5}{2})}{\sqrt{(5-\frac{7}{2})^2+(2-\frac{7}{2})^2}*\sqrt{(2-\frac{5}{2})^2+(3-\frac{5}{2})^2}} = ?\ [1\text{pt}]$$

Q6— Part 2. Calculate U4's predicted rating on I4 (rounded off to nearest 1 digits):

$$\frac{5(-1)+3(0)+1(-1)}{1+0+1} = \textbf{?}\ [2\text{ pt}]$$

1. (True false - 1pt) In BigCLAM, we define the probability of an edge Pa(u,v) between two nodes u and v in the community A as Pa(u,v) = 1 - exp(FuA*FvA) , where FuA is the membership strength of node u to community A.

2. (True False - 1pt) In a DAG, dividing nodes into two sets so that the cut is maximized is considered a good partition.
3.
4. [3%] For the given graph, generate the Adjacency Matrix, Degree Matrix, and Laplacian Matrix and answer the following:



	1	2	3	4	5	6

1		B				
2					F	
3	A		D			
4						
5				C		
6						E

a. [1%] For the Adjacency Matrix, fill in the values below:

   i. A+B = ?

   ii. B-C = ?

   iii. [1%] For the Degree Matrix, fill in the values below:

   iv. D = ?

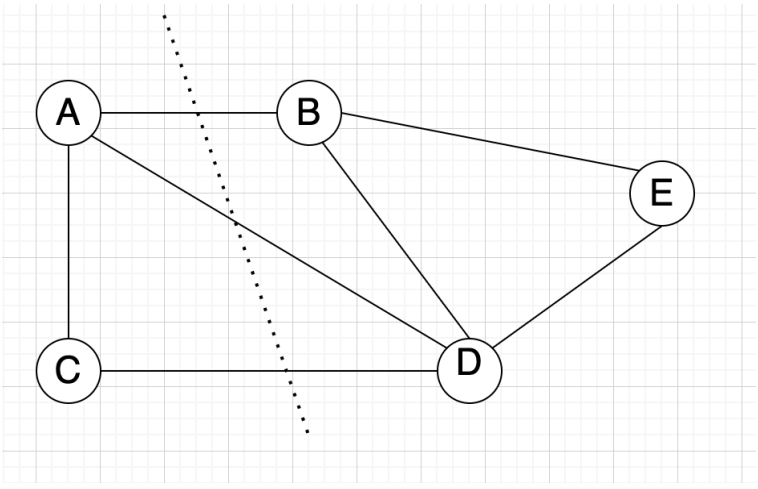
   v. E = ?

b. [1%] For the Laplacian Matrix, fill in the values below:

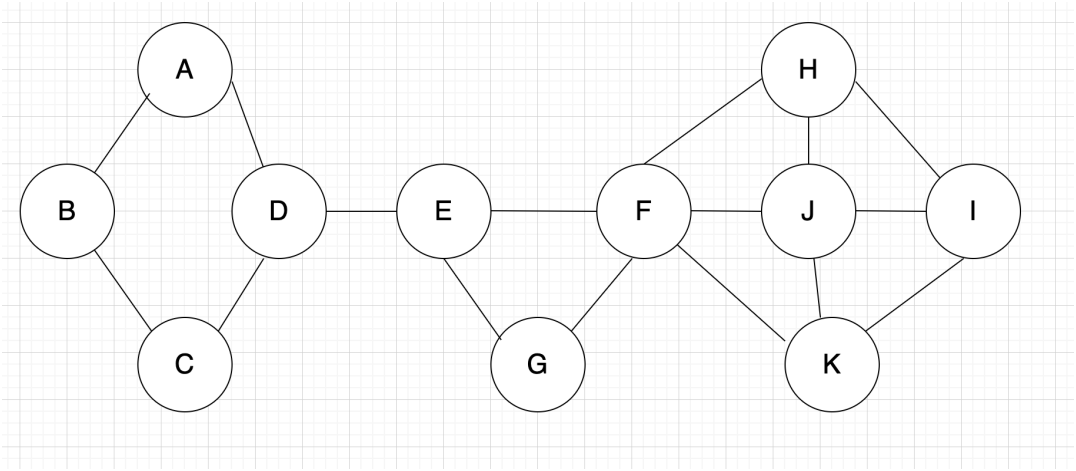
   i. C = ?

   ii. F = ?

5. [1%] Calculate the normalized cut for the given graph: (show answers to two decimal places)



6. [1%] Use the Girvan-Newman algorithm to perform the credit calculation starting from the following nodes:



   a. What is the final credit assigned to E when starting from E?
   b.
7. (True False - 1pt) In the Maximum Likelihood Estimation (MLE) model, the model has parameters for determining the probability of generating any instances of the artifact

8. (Multiple answer - 2pt) Select all the true statements:
   a. In AGM, if there are two nodes u and v in 2 communities C and D, the probability an edge exists between u and v is 1-(1-Pc)(1-Pd). (Both u and v are present in both C and D)
   b. In BigCLAM, if nodes u and v are not in the same community, Pa(u, v) = 0.
   c. AGM can express nested community structures
   d. According to the Affiliation-Graph model, if u and v share no communities, then P(u,v) = 0

Q1. (True or false 1pt)
[1%] When streams of data arrive at a rapid rate, it is practical to store them all in the main memory to answer queries.

Q2. (True or false 1pt)
[1%] In Fixed-size sampling, every new element always replaces an existing element which is picked uniformly at random.

Q3. (Multiple Answer 2pt)

[2%] Which of the following statements are true about Bloom Filtering
A. Bloom filtering can have false positives, but no false negatives.
B. Bloom Filters use hash functions to map elements to a bit array
C. Bloom Filtering is used to find the number of times an element appeared in a set.
D. In Bloom Filtering, assuming we have chosen an appropriate number of hash functions, The larger the bit array and the lesser the elements inserted, the lower will be the false positive rate

Q4. (Multiple Answer 2pt)
[2%] Which of the following is correct for DGIM Algorithm?
A. DGIM algorithm uses $O(\log^2 N)$ bits storage.
B. Rightmost of each bucket is always 1.
C. Size of older buckets is always greater than the size of the new ones.
D. DGIM estimates the number of 1's in the window with no more than a 50% error.

Q5. (Filling the Blanks - 1pt)
[1%] Consider the stream: 1 0 1 1 0 1 1 0 1 0, where new elements are added on the right.
According to the DGIM algorithm, the current state is: [1 0 1] [1 0 1] [1] 0 [1] 0.
What are the elements in the leftmost bucket after **another** bit of value 1 arrives and the stream becomes: 10110110101?

Your answer should look like *1110111* and have no spaces between them

Q6. (Fill in the blank - 1pt)
[1%] Consider a Bloom Filter implemented as follows:

- Initialize an 8-bit array B with each bit set to 0
- Two hash functions are being used:

h1(x) = (5*x + 13) % 8
h2(x) = (9*x + 7) % 8
Consider the stream (3, 11, 6). Now build the filter.

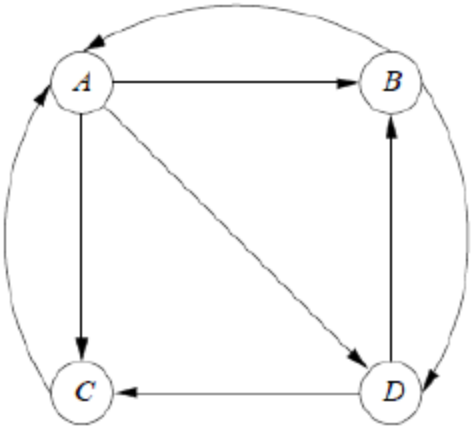
Your answer should look like *00110011* i.e., 8 digits <u>without any space</u>

Q7. (Multiple Choice - 1pt)
[1%] Given a stream S: a b c b d a c d a b d c a c b. Assuming the starting index of the stream is at 0, what is the estimated 2nd moment of S with two starting variables at position 3 and 7 ?
A. 60
B. 15
C. 50
D. 75

Q8. (Multiple Choice - 1 pt)
[1%] Suppose we apply the Flajolet-Martin algorithm with a single hash function h, to estimate the number of different elements in this stream of integers consisting of one 1, two 2's, three 3's, and so on, up to seven 7's. h(i) is simply written as a 32-bit binary number (e.g., h(1) = 00...001, h(2) = 00...010). What estimate does h give as the number of distinct elements in this stream?
A. 4
B. 16
C. 2
D. 8

1. The time complexity using Power Iteration to calculate the PageRank of a graph with n nodes is O(kn^2) (where k is the number of iterations) [1 point]
    a. True
    b. False
2. If page j with importance rj has n out-links, each link gets how many votes? [1 point]
    a. r_j / n
    b. $(r_j + 1) / n$
    c. R_j
    d. $r_j / (n + 1)$
3. The web can be considered as an undirected graph, in which nodes are web pages, edges are hyperlinks. [1 point]
    a. True
    b. False
4. Which of the following statements are true? [2 point]
    a. Spam pages are less connected so there is less chance to attract random surfer
    b. Page is more important if it has more outgoing links
    c. If a page is important, then random surfer can easily find it
    d. Page is not important if it attracts a large number of surfers
5. Gaussian Elimination is always more efficient than Power Iteration in all cases.[1 point]
    a. True
    b. False
6. Given the graph below. What is the transition matrix(M) for it?

Write your answer as a fraction (Example: 1/4) if the answer is a fractional value, else enter the integer value (Example: -1, 0, 1, etc.). Usage of Decimals is not NOT allowed. [2 points]



	A	B	C	D
A				
B				
C				
D				

7. Using power iteration to calculate the PageRank, what is v^1 for the graph in question 5 (v^0 is the initial vector)? [2 point]

V^1

Q1. According to the theory of Markov process, for graphs that satisfy certain conditions, no matter what the initial probability distribution at time t = 0, the stationary distribution _____ , _____.

Question 1 options:

A. might not be unique, but eventually will be reached

B. is always unique, but might not be reached

C. might not be unique, and might not be reached

D. is always unique, and eventually will be reached

Q2. Taxation enables teleporting.

Question 2 options:

A. True

B. False

Q3.One viable method to choose the topic set in Topic-Specific PageRank is to look at the words that appeared in the recently searched web pages queried by the user.

Question 3 options:

A. True

B. False

Q4. When we compute hubbiness and authority via mutual recursion, we start with authority = h vector of all 0's.

Question 4 options:

A. True

B. False

Q5. When dealing with PageRank dead-ends using the method of deleting all the dead-ends from a graph, several passes of prune and propagate will give approximate values for dead-ends by propagating values from the reduced graph.

Question 5 options:

A. True

B. False

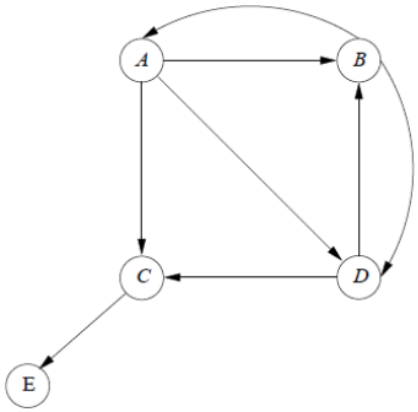
Q6. Consider PageRank with taxation ($\mathbf{v}' = \beta M\mathbf{v} + (1-\beta)\mathbf{e}/n$) which is usually used to deal with spider-traps.
When there is a dead-end in the graph, the sum of components in v' might be larger than the sum of components in v.

Question 6 options:

A. True

B. False

Q7. Which node(s) in the following graph has/have the highest authority score?



Question 7 options: (check box)

C

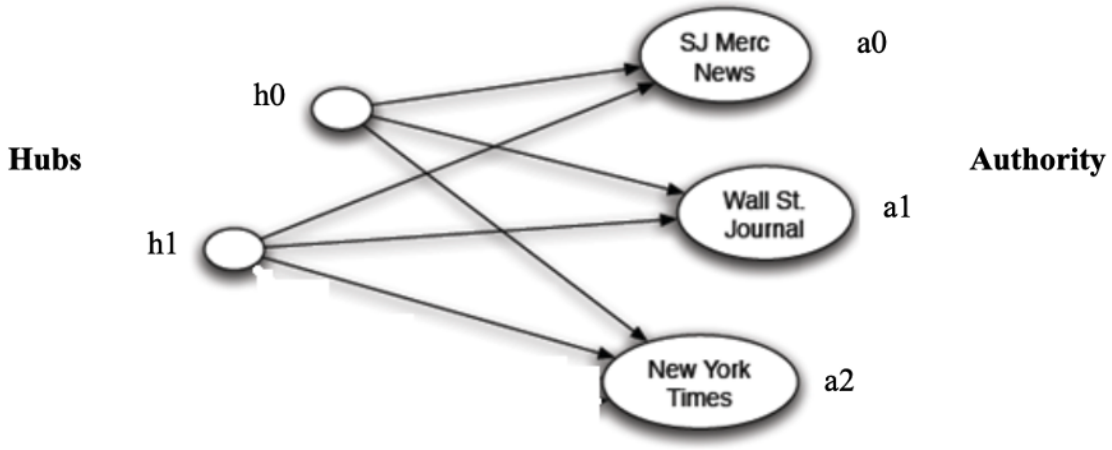
D

E

B

A

Q8.



Question 8 options:

Calculate Hub score(for h0 and h1) and Authority score(for a0, a1 and a2) for the following web graph (a0, a1, a2 and h0, h1) (Note: Calculate using Mutual Recursion till **convergence** only. Assume 1.0 for all initial scores.)

a0 = ?

a1 = ?

a2 = ?

h0 = ?

h1 = ?