Quiz 1 - Map Reduce

1. [1 point] Which of the following is NOT true for Combiners?

A. Combiners save network time by pre-aggregating values in the mapper
B. 1 Combiner combines the values of all keys from all the mappers (all nodes).
C. The Combiners instances are run on every node that runs the reduce tasks.
D. The Combiner is a “mini-reduce” process.

2. [1 point] **Select ALL** of the problems/data NOT 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 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] Transformations are lazy operation in Spark?

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

Quiz 2 - Frequent Itemsets I

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] What are the **item counts** produced in pass 1 and which of these items are frequent?

Item	Count	Frequent
a		
b		
c		
d		
e		
f		

1.2) [2 pts] For pass 2, **which are the candidate pairs** for each basket? (Only include the pairs that **will be counted**.)

`Basket	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
(a, b)		
(a, c)		
(a, d)		
(b, c)		
(b, d)		
(c, d)		

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?

Quiz 3 - Similar Sets

1. [2 Points] Answer the following questions:
(a) How many 2-shingles does **Humpty** have?
(b) How many 2-Shingles does **Dumtpy** have?
(c) What is the Jaccard Distance between the two? (Write the answer as a fraction. eg: 2/5)

2. [1 point] Consider the following characteristic matrix of two sets: S1 and S2.
Suppose we use the two hash functions: h1(x) = (x + 3) % 8 and h2(x) = (2 * x) % 5 to generate signatures of the sets as shown in the table below.

2.1 Fill in the blanks with the new row numbers generated by each hash function.

Row	**S_1**	**S_2**	**(x+3) %8**	**(2x) %5**
0	1	0		
1	1	1		
2	1	1		
3	0	1		
4	1	1		
5	1	0		
6	0	1		

3. **2.2** [2 points]Construct a signature for S_1 and S_2 based on the minhash values obtained from the $f_1(x)$ and $f_2(x)$ in Question 2.

Fill in the blanks with the values of the signature matrix:

	S_1	**S_2**
h_1		
h_2		

4. **2.3** [1 point] For the below questions, write your answer as a fraction. (eg: 1/3)

Using the signature matrix from Question 3:

(a) Estimate the Jaccard similarity of S_1 and S_2

(b) What is the actual Jaccard similarity of S_1 and S_2?

5. **2.4** [1 point] Using the Jaccard similarities calculated from Question 4, determine the following

Is the estimate close to the actual Jaccard similarity? If not, how can the estimate be improved?

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.

6. [2 points] Suppose that two sets are considered to be similar if their Jaccard similarity is greater than or equal to 0.6. Consider two sets S_1 and S_2. Suppose that their actual Jaccard similarity is 0.8. Consider their minhash signatures S_1' and S_2', each having 100 minhash values. Suppose the signatures are divided into 25 bands with 4 rows in each band. That is, b = 25, r = 4. Locality-sensitive hashing (LSH) is then applied to the signatures to obtain candidate pairs of sets. What is the probability that S_1 and S_2 are **not** identified as a candidate pair (i.e., false negative rate)?

a. 0.00019 %
b. 0.00137 %
c. 3.11 %
d. 96.88 %

7. [1 point] What is the effect of following on the false positive and false negative rate in LSH? Increasing bands *(b)*, keeping rows *(r)* constant

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

Quiz 4 - Recommendation System 1

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]

Q2. Select **ALL** of the statements that are TRUE about **Content-based Approach**. [1 pt]

A. It is able to recommend new & unpopular items
B. It faces cold-start or sparsity problems
C. It recommends items outside user’s content profile
D. It is unable to exploit quality judgments of other users

Q3.

	HP1	HP2	HP3	TW	SW1	SW2	SW4
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].

	HP1	HP2	HP3	TW	SW1	SW2	SW4
A							
C							

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

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

	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 =
W3,4 =

W3,5 =

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]

Quiz 5 - Recommendation System 2

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. [True/False] **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. [True/False] **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. [True/False] 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	**?**

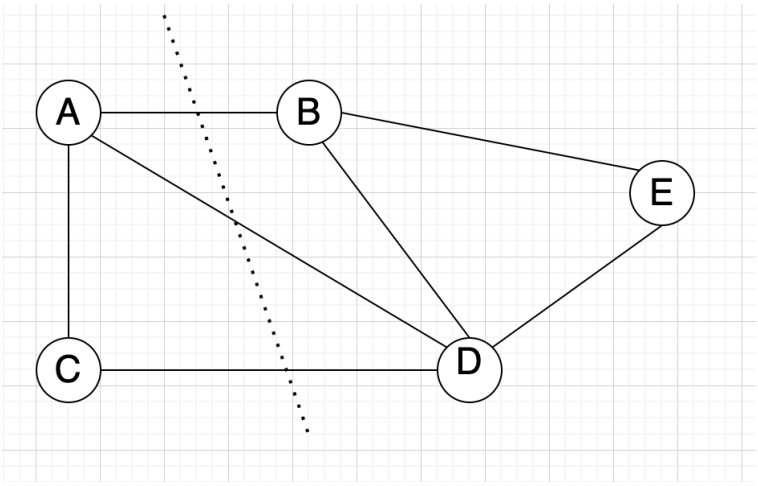
Q6— Part 1. Calculate weights **using average ratings of only co-rated items**. (rounded off to nearest 2 digits):

W1,4 =**[**1 pt**]**
W2,4 = [1 pt]
W3,4 = [1pt]

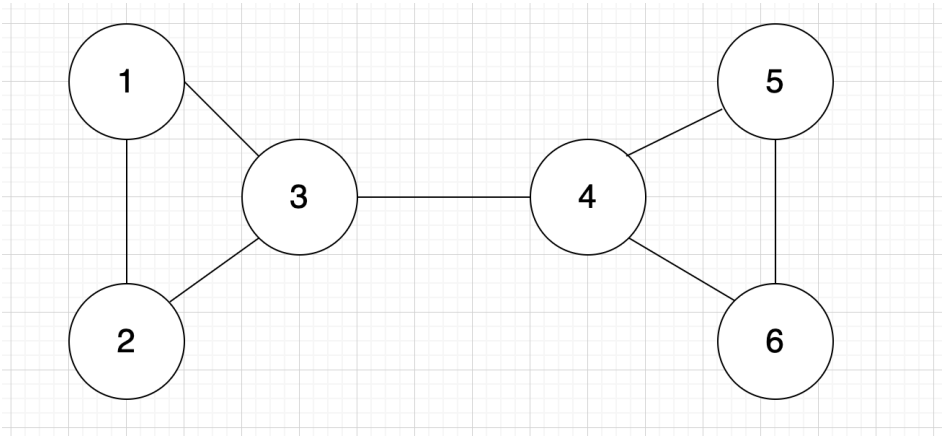
Q6— Part 2. Calculate U4's predicted rating on I4 (rounded off to nearest 1 digits): [2 pt]

Quiz 6 - Social Network 1

1. For each of the statements below, select True if the statement is always and unconditionally true, or False if it is always false, sometimes false, or just does not make sense:
    a. [1%] In betweenness, a low score is "good": suggests that edge (a,b) runs between two different communities
    b. [1%] In a DAG, dividing nodes into two sets so that the cut is maximized is considered a good partition.
    c. [1%] Given NxN symmetric matrix, eigenvalues are non-negative real numbers .
    d. [1%] The maximum number of edges in a 10-node undirected complete bipartite graph is 20.
2. [1%] Calculate the normalized cut for the given graph: (Give answers in decimal, do not leave in fraction)



3. [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			F			
2	A					
3		B				
4				C		
5					D	
6						E

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

      i. A+B =

ii. B+C =

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

i. D =

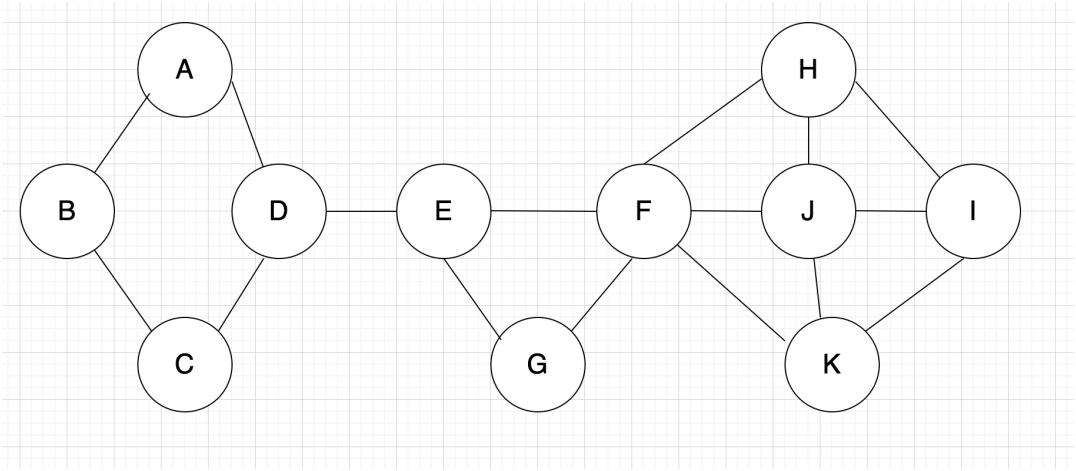
ii. E =

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

i. F =

ii. C =

4. [2%] Use the Girvan-Newman algorithm to perform the credit calculation starting from the following nodes:



a. [1%] What is the final credit assigned to E when starting from E?

b. [1%] What is the final credit assigned to G when starting from G?

Quiz 7 - Social Network 2

1. (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

2. (true false - 1 pt) In the Maximum Likelihood Estimation (MLE) model, the value of parameters that gives the smallest value of the likelihood is the correct model for the artifact

3. (true false - 1 pt) Affiliation Graph model uses non-binary memberships to generate the network.

4. (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 membership strengths for nodes u and v in community A are large numbers, Pa(u, v) is close to 0.
    c. In AGM, if there are two nodes u and v in 2 communities C and D, the probability of no edge existing between Nodes u and v is (1-Pc)(1-Pd). (Both u and v are present in both C and D)
    d. AGM can express nested community structures

5. (True false - 1 pt) According to the Affiliation-Graph model, if u and v share no communities, then P(u,v) = 0

6. (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.

7. (Multiple fill in answer - 3 pts) Given the following matrix:

Fu	0	0	0.7	0.8
Fv	1.1	0	0	1.2
Fw	1.3	1.8	1	0

Compute the probability P(n, m) of at least one common community link for the membership matrix above; show answers to two decimal places:

Quiz 8 - Data Streams

Q1. (True or false 1pt)
When dealing with static data, storing it in the main memory to answer queries is not a good idea.

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

Q3. (Multiple Answer 2pt)
Which of the following statements are not true about Bloom Filtering
A. Bloom filtering can have false negatives, but no false positives.
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)
Which of the following is correct for DGIM Algorithm?
A. DGIM algorithm uses $O(\log^2 N)$ bits storage.
B. Rightmost of each bucket can be either 0 or 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)
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. (Multiple Choice - 1pt)
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. 15
B. 50
C. 60
D. 75

Q7. (Multiple Choice - 1 pt)
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. 2
B. 4
C. 8
D. 16

Q8. (Fill in the blank - 1pt)
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 without any space

Quiz 9 – Clustering

Q1. [1 point] (True or false 1pt **< 1 min**)

K-means is a hierarchical clustering algorithm that utilizes centroid-based clustering, CURE is a density-based clustering algorithm that uses representative points, and BFR is a distance-based algorithm that handles large datasets with a focus on minimizing memory usage and I/O cost.

Q2 [1 point] **(Multiple Choice < 1 min)**

Given the strings "kitten" and "sitting", compute the edit distance and select the correct explanation of the calculation.

A) Edit distance: 3; Substitute 'k' with 's', insert 't', and delete 'e'.

B) Edit distance: 4; Substitute 'k' with 's', insert 'i', insert 't', and delete 'e'.

C) Edit distance: 3; Substitute 'k' with 's', substitute 'e' with 'i', and insert 'g'.

D) Edit distance: 5; Substitute 'k' with 's', insert 'i', insert 't', substitute 'e' with 'i', and substitute 'n' with 'g'.

Q3 [1 point] **(Multiple Choice < 1 min)**
Which of the following statements are true regarding the BFR (Bradley-Fayyad-Reina) algorithm and its requirements? (Select all that apply)

A) BFR algorithm is specifically designed for datasets with low-dimensional feature spaces.
B) BFR algorithm assumes that the clusters need to be normally distributed around a centroid in a Euclidean space.
C) BFR algorithm requires prior knowledge of the distributions on each dimension.
D) BFR algorithm can handle datasets that do not fit in the main memory.

Q4 [1 point] **(Multiple Choice < 2 min)**
We are given a set of points named by their (x, y) coordinates. Initially, each point is in a cluster by itself and is the centroid of that cluster. Now we want to combine the points into clusters using L_2 norm (Euclidean Distance). Which points are the first to be merged:

A. (4, 10), (4, 8)
B. (3, 4), (2, 2)
C. (7, 10), (6, 8)
D. (11, 4), (12, 3)

Q5 [2 points] **(Filling the blanks < 3 min)**

Suppose a cluster consists of points (5,1), (6, -2), and (7,0). The representation of this cluster as in the BFR algorithm is represented as N, SUM, and SUMSQ. Compute the variance of the cluster in each of the two dimensions. (round to 3 decimals)

First Dimension:
Second Dimension:

Q6 [2 points] **(Filling the blanks < 4 mins)** Perform a hierarchical clustering of the one-dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, and 81, assuming the clusters are represented by the centroid (average), and at each step, the clusters with the closest centroids are merged, continue until only two clusters remain.
What is the centroid of the cluster having more members:

Q7. [2 points] (**Filling the blanks < 3 mins**) The SSE (sum squared error) is the sum of the squares of the distances between each of the points of the cluster and the centroid. Given a set of data points and their cluster assignments, compute the Sum of Squared Error (SSE) for each clustering scenario and select the one with the lowest SSE.

Data points: (1, 2), (3, 4), (5, 6), (7, 8)
We want to cluster 4 points into 2 clusters, with 2 points in each cluster. What is the minimum SSE:

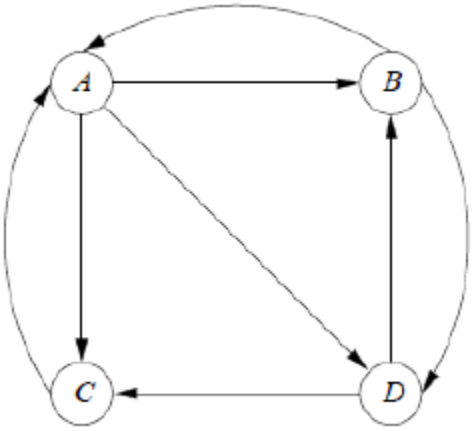
Quiz 10 - Link Analysis 1

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

Quiz 11 - Link Analysis 2

1. (1 pt, <3min) 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 _____ , _____.
    a. is always unique, and eventually will be reached
    b. might not be unique, but eventually will be reached
    c. is always unique, but might not be reached
    d. might not be unique, and might not be reached

2. (1 pt, <1min) Teleporting enables taxation. [True/False]

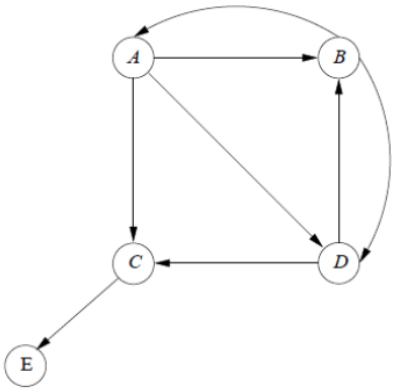
3. (1pt, <1min) One viable method to choose the topic set in Topic-Specific PageRank is to look at the words that appeared in the recently searched web pages queried by the user. [True/False]

4. (1 pt, <1min) When we compute hubbiness and authority via mutual recursion, we start with authority = h vector of all 0's. [True/False]

5. (1 pt, <1min) 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. [True/False]

6. (1 pt, <1min) 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 smaller than the sum of components in **v**. [True/False]

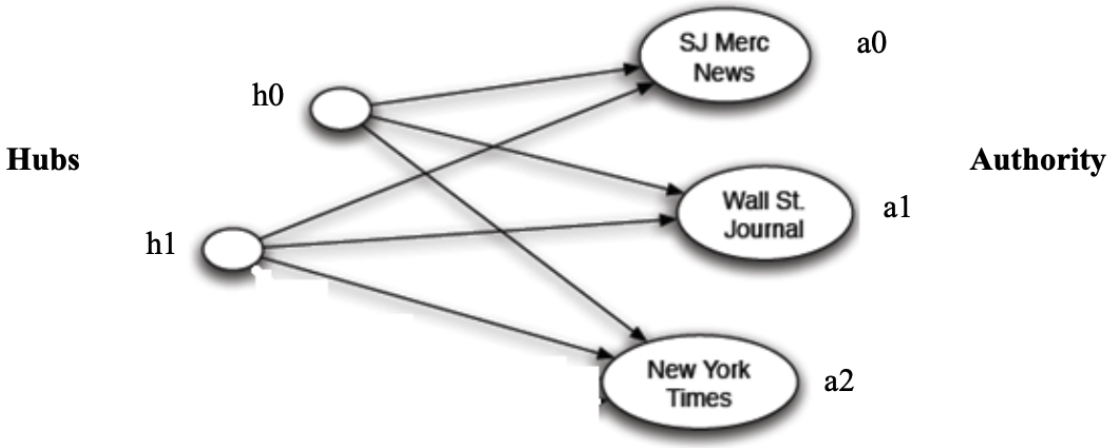
7. (1.5pt, <3min) Which node(s) in the following graph has/have the **highest authority** score?



    a. A
    b. B
    c. C

d. D
e. E

8. (2.5 pts) 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.)**



Notice: As for the last lecture on Web Advertising, since we don't have any quiz questions about it, please go over the slide on your own. LOL. GOOD LUCK!