Question 1 - Graph Analysis and PageRank [24 Points]

a) How can we model the internet using graphs? [2 points]

[1 point] Mentions that webpages are vertices
[1 point] Mentions that links are edges

b) Why is it important to model the internet using graphs? [2 points]

This internet modeling can be used to compute the importance of each web page in the network

c) Define PageRank [2 points]

PageRank is a number/algorithm used to define the importance of each webpage.

d) List **two** problems with the basic PageRank algorithm. In at most one sentence, describe the solution to these problems discussed in class. [6 points]
[2 points per problem][2 points for solution]

Problems : Dead End→ some pages has no out link
Spider Traps → all outlinks are within the group

Teleports solve the above problems.
Dead End: Never get stuck in a spider trap by teleporting out of it in a finite number of steps
Sprider tarps: Make matrix column stochastic by always teleporting when there is nowhere else to go

e) State one challenge to processing a graph in MapReduce, as well as a possible solution. [4 points]

Following are the challenges with possible solutions:
[2 point for mentioning one challenge]
[2 points for mentioning the possible solution]
Any One challenge with one possible solutions is OK]

1. **Challenges: How do we represent our graph?**

Solution : Our graphs will be represented as a collection of `Node` objects

```
Node: nodeId,
      distanceLabel,
      adjancencyList[nodeId, distance],
      ...
```

Input the graph as text and parse it to build our `<key, value>` pairs

So what are our `<key, value>` pairs?

2. Challenge: What are our <key,value> pairs?

Solution : From mapper to reducer two types of `<key, value>` pair:

```
<nodeId n, Node N>   // nodeId to Node object
<nodeId n, distance> // nodeId to distance so far
```

3.How do we iterate through various stages of processing?

Solution: Each *iteration* in the algorithm is a MapReduce job.

Iterations and termination are coordinate by an external *driver* application
The first iteration starts at the source node (with distance 0)
It updates and emits all distances for nodes in the adjacency list
The next iteration takes the output from the previous and updates/emits all distances for nodes connected to this set of nodes and Continue until termination

4.When/how do we terminate execution?

Solution: Terminate when the graph has reached a steady state:

- **All the nodes have been labeled with min distance**
- **Labels no longer change between iterations**

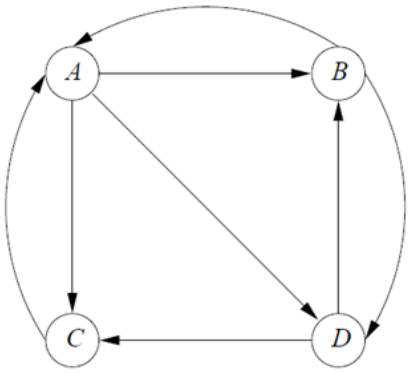


Fig 1: Example Graph

Figure 1 is an example of a tiny version of the Web, with only four pages. Page ***A*** has links to each of the other three pages; page ***B*** has links to ***A*** and ***D***; page ***C*** has a link to ***A***; and page ***D*** has links to ***B*** and ***C***.

f) Write the flow equations for the graph in Figure 1. [4 points]

ra=rb/2 +rc
rb=ra/3 + rd/2
rc=ra/3 +rd/2
rd=ra/3 +rb/2

g) Generate the stochastic adjacency matrix, **M**, for the graph in Figure 1. [4 points]

	A	**B**	**C**	**D**
A	0	½	1	0
B	⅓	0	0	½
C	⅓	0	0	½
D	⅓	½	0	0

Question 2 - Bias [20 Points]

For each of the following scenarios, state one of the six types of bias discussed in class that may apply to the specific scenario. In 2-3 sentences, describe **how that specific bias is manifesting** in the scenario, and **propose a fix**.

FOR EACH SCENARIO
[1 point] For answer
[2 points] Valid explanation
[2 points] Valid fix

a) Imagine that we have built and deployed a spam filter in the CSE department using what we've learned in class. However, after deploying the spam filter, we've received some complaints from faculty that they are still receiving almost as many spam emails as they were before. When we look into these complaints, we realize that these complaints are mostly coming from female faculty members. **Note that female faculty make up around 16% of the total faculty in the CSE department.** [5 points]

Representation Bias: Female faculty are underrepresented in the dataset, get more data on female faculty or sample differently
Evaluation Bias: A single evaluation metric is used which works well for male faculty but ignores inaccuracy in female faculty. Use multiple evaluation metrics
Aggregation Bias: A single model is applied to both populations but does not work well on female population. Train separate models for each group.

Other answers are possible as long as the explanation is good.

b) Imagine UB has started an initiative to better track and analyze student attendance habits by utilizing wearable technology. They have developed an app that students can install on smart watches, that tracks how frequently students attend class. Using this information UB hopes to get more data on the relationship between attendance and GPA. [5 points]

Representation Bias: Part of the population that does not own smart watches are not represented. Use different data collection mechanisms.

Other answers are possible as long as the explanation is good.

c) Imagine that your favorite sports team has decided to turn to data intensive computing in order to improve their roster building decisions. To build the best team, they want to figure out which players are the **most skilled**. In order to do this, they collect data on **how many games each player has won** over the past 5 years. [5 points]

Measurement Bias: This solution conflates wins with player skill. Use more features more directly related to player skill.

Other answers are possible as long as the explanation is good.

d) Home assistants like Google Home and Amazon Alexa are becoming more and more prevalent as the technology improves. Some figures claim that Google Home can recognize human speech with up to 95% accuracy! However, other reports cite much lower accuracy for female users. [5 points]

Representation Bias: Females may be underrepresented in the dataset used to train. Get more data on female voices.
Evaluation Bias: A single evaluation metric is used which works well for male users but ignores inaccuracy in female users. Use multiple evaluation metrics
Aggregation Bias: A single model is applied to both populations but does not work well on female population. Train separate models for each group.

Other answers are possible as long as the explanation is good.

Question 3 - Spark [22 Points]

a) Narrow dependencies: [5 points]

i) Define what it means for a dependency to be narrow

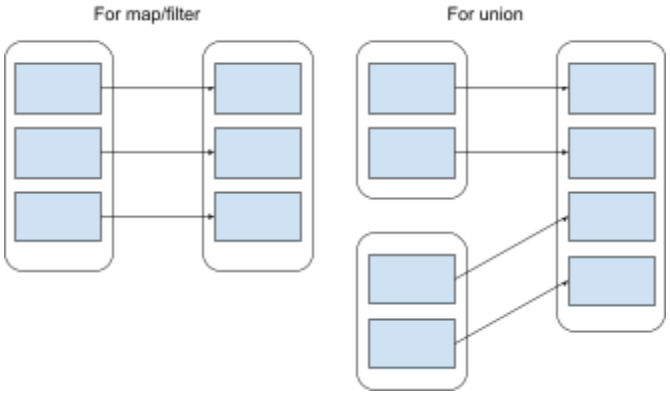
[2 points] A dependency where each child depends on only one parent

ii) Give an example of a narrow dependency

[1 point] Map, filter, union, etc

iii) Draw the diagram of the given example

[2 points] Matches one of below (based on ii)



b) Wide dependencies: [5 points]

i) Define what it means for a dependency to be wide

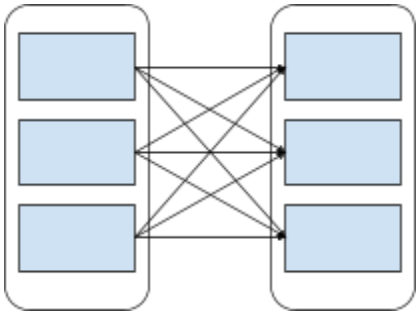
[2 points] A dependency where each child depends on multiple parents

ii) Give an example of a wide dependency

[1 point] groupByKey, reduceByKey

iii) Draw the diagram of the given example

[2 points] Matches below



c) Give one reason why narrow dependencies are less expensive than wide. [4 points]
[4 points for valid reason]
- **Wide requires communication to shuffle data**
- **Narrow can be pipelined**
- **Narrow are cheaper to recompute if fault occurs**

d) Why do we need a separate framework for streaming data? [4 points]

Continuous flow of data/event/streaming data Requires large clusters to handle workloads and Requires latencies of a few seconds . Existing frameworks can not do both.

Example: Tweet event , stream of log messages

e) State one difference between a stateless and a stateful operation in Spark Streaming. [4 points]

One difference is Ok

Stateless operations they know nothing about any previous batches.
Stateful operations have a dependency on previous batches of data

Question 4 - Spark [22 Points]

For Question 4, refer to the following Spark code. Assume that the number of files in the **`files`** list is 2, and that **`fullOuterJoin`** is a transformation resulting in a narrow dependency.

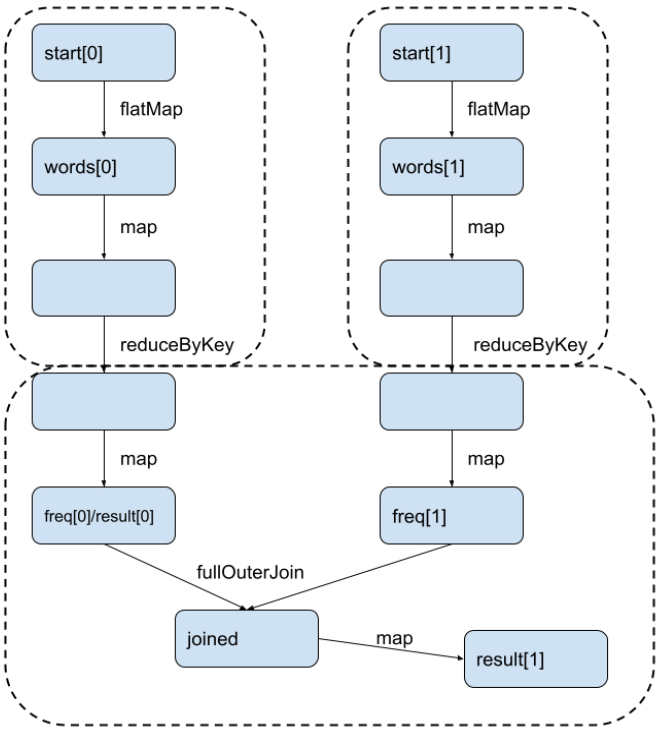
```
1  result = None
2  for file in files:
3      start = sc.textFile(file)
4      words = start.flatMap(lambda x: x.split(' '))
5      N = words.count()
6      freq = words.map(lambda x: (x, (1, file))) \
7                  .reduceByKey(lambda a,b: (a[0] + b[0], a[1]))
8                  .map(lambda x: (x[0], (x[1][0] / N, x[1][1]))
9
10     if result == None:
11         result = freq
12     else:
13         joined = result.fullOuterJoin(freq)
14         result = joined.map(lambda x: (x[0], maxFreq(x[1]))
15 result.collect()
```

a) Draw the lineage graph for the RDD **`result`** right before the call to **`result.collect()`**. Include nodes for all intermediate RDDs, even if they are unnamed. [10 points]

[10 points if perfect (ignoring boundaries)]

Subtract 1 for:

- **each missing/extra RDD**
- **each missing/extra arrow**



b) How many stages will your DAG be broken up into for execution? Draw the stage boundaries in your diagram. [10 points]

[4 points] 3 stages
[6 points] Correct boundaries (shown above)
[Half Credit] Off by one with reasonable boundaries

c) How many "jobs" will the above code run? [2 points]
[2 points] 3 jobs
[1 point] 2 jobs (off by one, usually only count one count())

Question 5 - Hive/Cloud [22 Points]

a) Lists two advantages of HIVE [4 points]

HiVE is Useful for people who aren't from a programming background as it eliminates the need to write complex MapReduce program.

Hive supports any client application written in Java, PHP, Python, C++ or Ruby by exposing its Thrift server

b) In what situation would you want to use HIVE [4 points]

[Any 4 situations are OK]

Data Mining
Log Processing
Document Indexing
Customer Facing Business Intelligence
Predictive Modeling

c) Name two specific challenges that cloud-based applications must be able to handle effectively. [4 points]

[2 points per correct challenge]

- **Load balancing**
- **Fault tolerance**
- **Multi-tenancy**
- **Elasticity**
- **Scalability**

d) Name the 3 cloud service models discussed in class in order of how much the underlying resources are abstracted. Put the service model with the highest amount of abstraction first, and the model with the lowest amount last. [6 points]

[1 point each for mentioning service models]
[3 points if also in correct order]
[1 point if in reverse order]
SaaS (Software as a Service), PaaS (Platform as a Service), IaaS (Infrastructure as a Service)

e) Give one concrete reason why Spark might perform better than MapReduce for iterative applications specifically. [4 points]

[4 points] In-memory caching, keeps data in memory rather than disk

Question 6 - Misc Review [15 Points]

In a MapReduce application computing relative word co-occurrence the mappers output the following key-value pairs:

((dog, tail), 4), ((dog, spot), 1), ((pass, fail), 8), ((dog,fail), 1), ((pass, go),2), ((dog,go),7), ((pass,dog),1), ((fail, class), 6), ((fail, pass), 3), ((fail, epic), 10)

a) Show how these key-value pairs might be **sorted** and **partitioned** during the barrier in between map and reduce in order to be able to compute the relative co-occurrence. Write out the key value pairs in a valid order, and draw a line between two of the key value pairs to demarcate how they could be partitioned across **two reducers**. [4 points]

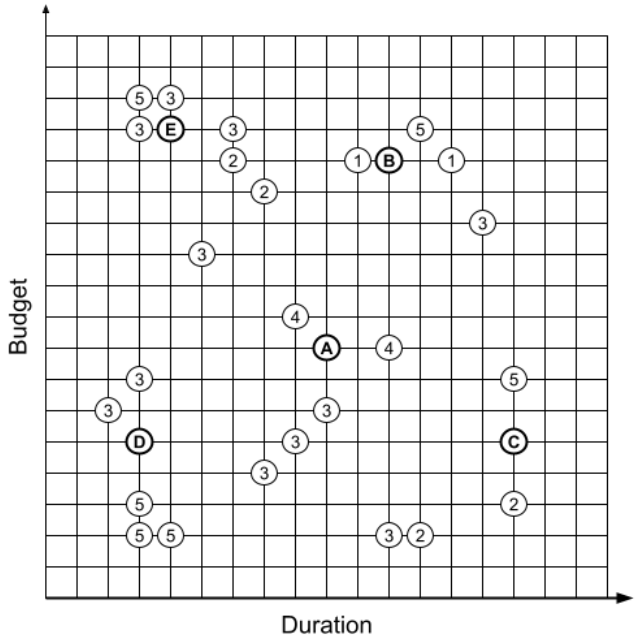
[2 points for a valid sorting – all (dog,*), (pass,*) and (fail,*) contiguous (any order)]
[2 points for a valid partitioning – don't split up any (dog, *), (pass, *), (fail,*)]

Example:
((dog, tail), 4), ((dog, spot), 1), ((dog,fail), 1), ((dog, go), 7) | ((pass, fail), 8), ((pass,go),2), ((pass,dog),1), ((fail, class),6), ((fail, pass),3), ((fail,epic),10)

b) If we want to apply order inversion to more efficiently compute the relative word co-occurrence, what additional key-value pairs must be emitted by the mappers **(write them out explicitly),** and where do those key-value pairs need to be placed in the sorted order from part (a)? [4 points]

[2 points for correct missing key-value pairs]
[2 points for stating / showing where they go]
((dog, *), 13) – Placed BEFORE first dog key
((pass, *), 11) – Placed BEFORE first pass key
((fail,*), 19) – Placed BEFORE first fail key

Below is a plot of movie ratings (1-5 stars) based on movie duration and budget. There are five movies (A, B, C, D, E) that are not rated. For the following questions you can assume the plot is drawn to scale.



c) We want to use k-NN to classify the unknown movies. Fill in the following table based on the output of k-NN when k=3 and k=5. In each box write the rating that the model would give each of the unknown points. [5 points]
[0.5 point per correct entry]

	A	B	C	D	E
k = 3	**4**	**1**	**2**	**3**	**3**
k = 5	**3**	**1**	**2**	**5**	**3**
Actual	**1**	**1**	**2**	**5**	**3**

d) Given the actual ratings for the movies shown in the table, what is the accuracy of our model when k = 3? When k = 5? [2 points]

[1 point] k=3: 60%
[1 point] k=5: 80%

Question 7 - Extra Credit [5 Points]

a) Name the 4 Vs of big data [4 points]

[1 point per]

Volume Velocity Variety Veracity

b) Name one reason why DIC is becoming more accessible [1 points]

- **Cloud computing**
- **Open source libraries/algorithms**
- **Other reasonable answer**

Scrap Paper