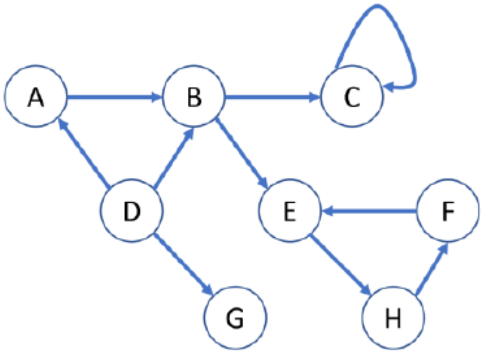
CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 4: Test Practice

1. Given the following graph,

1) how many dead ends are there in the graph? For each dead end (if any), please indicate the set of vertices forming the dead end.

2) how many spider traps are there in the graph? For each spider trap (if any), please indicate the set of vertices forming the spider trap.

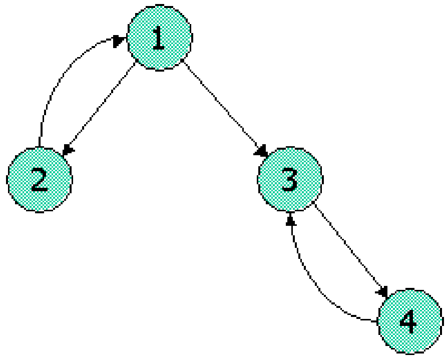


Answer:

2. True/False: In Topic-specific PageRank, random walker will teleport to any page with equal probability.

Answer:

3. Consider the following link topology.



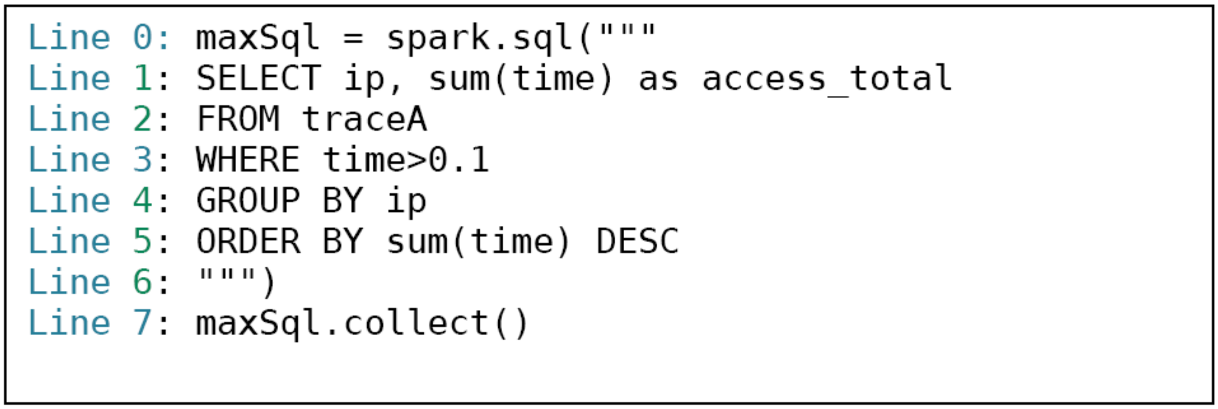
Write down the Topic-Specific PageRank equations for the following link topology. Assume that pages selected for the teleport set are nodes 1 and 2 (where teleports go to either node with equal probability). Assume further that the teleport probability, (1 - β), is 0.3.

Answer:

4. Show pseudocode for the compute() function for the PageRank with teleport (β = 0.85) over vertices algorithm in Pregel / Giraph. Set the initial PageRank value as 1/N (N is the number of vertices), Run 30 iterations and then stop. You can (if you choose) use the functions: getValue(), setValue(), getNumVertices(), getSuperStep(), getOutEdgeIterator().

Answer:

5. On a certain large Spark cluster, Rose creates a data frame named **traceA**, and writes the shown program to process the trace. The **traceA** data frame keeps the logs, and each log entry represents the log information of one web page access, including various fields: **ip**, the IP address of the log entry and **time**, the amount of time of that access.



```
Line 0: maxSql = spark.sql("""
Line 1: SELECT ip, sum(time) as access_total
Line 2: FROM traceA
Line 3: WHERE time>0.1
Line 4: GROUP BY ip
Line 5: ORDER BY sum(time) DESC
Line 6: """)
Line 7: maxSql.collect()
```

For the above codes, what are the potential performance bottlenecks? Please identify which lines cause the bottleneck and justify your answer.

Answer: