CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

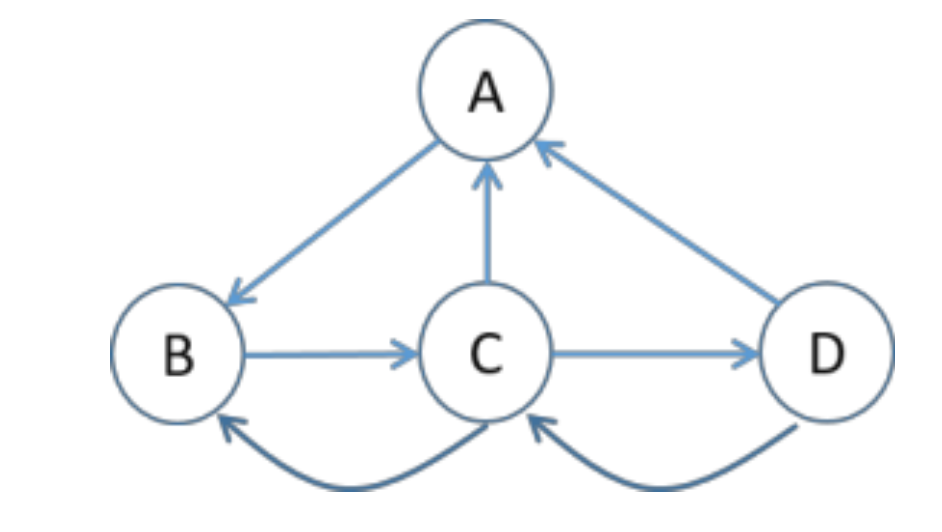
Tutorial 4: Test Practice

**1**. True/False: Personalized PageRank will generate the same ranking for any set of web pages.

**2**. True/False: In the Spark runtime, RDD cannot reside in the hard disk.

**3**. Set up the topic-specific PageRank equations for this graph, with teleport set {A} and β = 0.8 (jump probability = 1−β).

Note: you do not need to solve the equations for r(x).



**4**. In HDFS, each chunk is replicated for three times by default. In contrast, in Spark, RDD uses lineage for reliability. What are the major problems if Spark also uses replications for reliability?

**5**. Describe a system issue caused by skewed (i.e. highly imbalanced) data distributions on the following systems.

**Pregel**: computing PageRank on a 1 billion vertex graph, where the degrees of nodes are highly skewed (i.e. some nodes with very large degree)

**MapReduce**: computing word count, when the frequencies of different keywords is highly skewed (some keywords appearing many times)

**6**. Show pseudocode for the compute() function for the PageRank over vertices algorithm in Pregel / Giraph. You can (if you choose) use the functions: getValue(), setValue(), getNumVertices(), getSuperStep(), getOutEdgeIterator().