Exam TIES438 Big Data Engineering
Date: 15 December 2017

Name:_____

Each of the following questions is graded with 10 points. To pass, you need to obtain 50% of the points. Hand in all used papers.

1. How does min-hash work?

2. How can I maintain a uniform sample of a data stream which contains roughly 10% of the items seen?

3. I use a bloom filter ans created with a certain size and the optimal number of hash functions for inserting n items. Now, I insert more items into the bloom filter as I expected (more than n). What problems can this cause and why?

4. Why is the chosen distance metric important for a clustering algorithm?

5. Explain three of the ACID properties (atomicity, consistency, isolation, durability).

6. Amdahls law states that the maximum speed-up is $S(n) \leq \dfrac{1}{f + \dfrac{(1-f)}{n}}$. What does that mean?

7. What is a combiner in map-reduce?

8. Compute the PageRank for each node in the following graph: