CS 549—Distributed Systems and Cloud Computing
Assignment Five—Page Rank in Hadoop

1. Implementing a Hadoop task for PageRank (40%)

In addition to your code, you should demonstrate PageRank running on the simple graph in the assignment spec.

2. Running PageRank on Wikipedia data (30%)

Provide a video showing start-up of the task (which should run for hours).

3. Joining graph with node name information (10%)

Outputting node names (using an equipartitioned join).

4. Experimenting with different numbers of reducers (20%)

Experiment with different numbers of reducers and track how this choice affects the overall speed of your MapReduce job flow. In your report, include the details about what combinations of machines and reducers you used and your results. Also include your thoughts on why you saw these results (e.g. after a point, adding reducers may slow overall time, or running twice with the exact same arguments may show a notable difference).