

## Homework 6

**Student Number:**

**Name:**

**Problem 1.** (20 points) Two web search engines A and B each generate a large number of pages uniformly at random from their indexes. 35% of A's pages are present in B's index, while 55% of B's pages are present in A's index. What is the number of pages in A's index relative to B's?

**Problem 2.** (10 points) Why is it better to partition hosts (rather than individual URLs) between the nodes of a distributed crawl system?

**Problem 3.** (20 points) Instead of using the process depicted in shingle sketches, consider instead the following process for estimating the Jaccard coefficient of the overlap between two sets S1 and S2. We pick a random subset of the elements of the universe from which S1 and S2 are drawn; this corresponds to picking a random subset of the rows of the matrix A in the proof. We exhaustively compute the Jaccard coefficient of these random subsets. Why is this estimate an unbiased estimator of the Jaccard coefficient for S1 and S2?

**Problem 4.** (10 points) Why should the host splitter precede the Duplicate URL Eliminator?

**Problem 5.** (40 points) Web search engines A and B each crawl a random subset of the same size of the Web. Some of the pages crawled are duplicates – exact textual copies of each other at different URLs. Assume that duplicates are distributed uniformly amongst the pages crawled by A and B. Further, assume that a duplicate is a page that has exactly two copies – no pages have more than two copies. A indexes pages without duplicate elimination whereas B indexes only one copy of each duplicate page. The two random subsets have the same size before duplicate elimination. If, 45% of A's indexed URLs are present in B's index, while 50% of B's indexed URLs are present in A's index, what fraction of the Web consists of pages that do not have a duplicate?