Name: _____

CSE 4502/5717 Big Data Analytics

Exam I; October 12, 2021

Note: You are supposed to give proofs to the time and processor bounds of your algorithms. Read the questions carefully before attempting to solve them.

1. Input two sorted arrays A and B with $|A| = |B| = n$. It is known that $A \cap B = n^{3/4}$. The problem is to output an element that occurs in both A and B.

   (a) (12 points) Present a Las Vegas algorithm to solve this problem in $\widetilde{O}(n^{1/4} \log^2 n)$ time. Prove the run time of your algorithm. (Fact: $(1-x)^{1/x} \leq \frac{1}{e}$ for any $0 < x < 1$).

   (b) (5 points) How fast can you solve this problem deterministically?

2. (17 points) Input are k sequences $X_1, X_2, \ldots, X_k$ with n elements in each. The problem is to output an element y such that $y \in X_i$ for some i, $1 \leq i \leq k$ and y is greater than or equal to the median of X_i, for each i, $1 \leq i \leq k$. Present a Monte Carlo algorithm for this problem that runs in $O\left(k(\log n + \log k)\right)$ time. Show that the output of your algorithm is correct with a high probability.

3. (17 points) Input are two (not necessarily sorted) sequences A and B with n elements in each. The elements of A and B are integers in the range $[1, 10n]$. The problem is to check if they have a common element or not. Show that this can be done in $O(1)$ time using n common CRCW PRAM processors.

4. (17 points) Input are two sequences A and B, with $|A| = n$ and $|B| = m$. It is known that $m < n$. A is in sorted order and B is not. The problem is to compute $A \cap B$. Show how to do this in $O(\log n)$ time using m CREW PRAM processors.

5. (16 points) Input is a sequence X of n keys residing in a single disk. Each key in X is an integer in the range $[1, C]$, where C is a constant. Show how to sort X in two read passes (i.e., $\frac{2n}{B}$ read I/O operations, B being the block size) through the data.

6. (16 points) Input is a sequence X of n arbitrary real numbers residing in a single disk. The problem is to identify the least $\frac{M}{10}$ elements of X, M being the size of the core memory. Show that this can be done in one read pass (i.e., $\frac{n}{B}$ read I/O operations, B being the block size) through the data.