CSE 4502/5717 Big Data Analytics

Homework 1, due on September 27, 2022 at 3:30 PM

1. Input is a sorted array $a[1:n]$ of arbitrary real numbers. The array could only be of one of the following two types: 1) **Type I:** All the elements in the array are distinct; or 2) **Type II:** The array has $\sqrt{n}$ copies of one element, the other elements being distinct. Present a Monte Carlo algorithm that determines the type of the array in $O(\sqrt{n}\log n)$ time. Show that the output of your algorithm will be correct with high probability.(**Fact:** $(1-x)^{1/x} \leq 1/e$ for any $1 > x > 0$.)

2. $\mathcal{A}$ is a Monte Carlo algorithm for solving a problem Π that has a run time of $T_1(n)$ on any input of size n. The output of this algorithm will be correct with a probability of c, where c is a constant > 0. $\mathcal{B}$ is an algorithm that can check if the output from $\mathcal{A}$ is correct or not in $T_2(n)$ time. Show how to use $\mathcal{A}$ and $\mathcal{B}$ to create a Las Vegas algorithm to solve Π whose run time is $\widetilde{O}\left((T_1(n)+T_2(n))\log n\right)$.

3. Show that the maximum of n given elements can be found in $O(1)$ time using $n^{1+\epsilon}$ common CRCW PRAM processors, where ϵ is any constant > 0.

4. Present an $O(\sqrt{n})$ time algorithm for the selection problem. You can use up to $\sqrt{n}$ CREW PRAM processors.

5. Prove the following Lemma (known as the slow-down Lemma): If $\mathcal{A}$ is a parallel algorithm that uses P PRAM processors and runs in T time, then $\mathcal{A}$ can run on a P'-processor machine to get a run time of T' such that $T' = O\left(\frac{PT}{P'}\right)$, for any $P' \leq P$.

6. What happens to the I/O complexity of the sorting algorithm we discussed in class if we choose k to be $\frac{cM}{B}$ for some integer $c > 1$?

7. Present an efficient implementation of Djikstra's algorithm for the single source shortest paths problem on an out-of-core computing model with a single disk. What is the I/O complexity of your implementation?