Name: _____

CSE 5717 Big Data Analytics

Fall 2022 Exam IV

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. (17 points) X is a sequence of n arbitrary real numbers. The problem is to identify an approximate median of X. Specifically, we want to identify an element $x \in X$ such that $\left(\frac{n}{2} - a\alpha n^{2/3}\right) \leq rank(x, X) \leq \left(\frac{n}{2} + b\alpha n^{2/3}\right)$, with a probability of $\geq (1 - n^{-\alpha})$, for some constants a and b. Present an $O(n^{2/3} \log n)$ time algorithm for this problem. Prove the correctness of your algorithm.

2. (17 points) A sequence $X = k_1, k_2, \ldots, k_n$ is residing in a single disk. Each k_i is an integer in the range $[1, R]$, for $1 \leq i \leq n$. Show how to sort X in $O\left(\frac{n}{B}\frac{\log R}{\log(M/B)}\right)$ I/O operations.

3. (16 points) Input are a string S of length n and an integer $k < n$. The problem is to find a k-mer of S that occurs the largest number of times in S. Present an $O(n)$ time algorithm to solve this problem. For example, if $S = aabbbabaababa$ and $k = 2$, then one possible answer is ab since it occurs 4 times. ba also occurs 4 times. No other 2-mer occurs these many times.

4. (16 points) Let D be a database with n transactions from a set $I = \{i_1, i_2, \ldots, i_d\}$ of items. It is known that each transaction in D has $\leq c$ items, where c is a constant. Input are two thresholds *minSupport* and *minConfidence* for the minimum support and minimum confidence, respectively. Show that the total number of possible association rules whose support is $\geq$ *minSupport* and confidence is $\geq$*minConfidence* is $O(n)$.

5. (17 points) Present an $O(n \log n)$ time algorithm to compute $f(x) = \Pi_{i=1}^{\log n}(x + a_i)^{2^i}$, where $a_1, a_2, \ldots, a_{\log n}$ are scalars. The coefficients of $f(x)$ should be output.

6. (17 points) Consider a neural network with L layers. There are n neurons at each layer. Show that one forward propagation can be completed in $O(L \log n)$ time using $\frac{n^2}{\log n}$ CREW PRAM processors.