Name: _____

CSE 4502/5717 Big Data Analytics
Fall 2022 Model 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. Input is an array $A[1:n]$ of real numbers. There are 5 elements in this array that have $\frac{n}{10}$ copies each. The other elements occur exactly once each. The problem is to output any one of these 5 elements. Present an $\widetilde{O}(\log n)$ time Las Vegas algorithm to solve this problem.

2. Input are two $n \times n$ Boolean matrices A and B. The problem is to multiply these and output a Boolean matrix. Show that this problem can be solved in $O(1)$ time using n^3 common CRCW PRAM processors. For example, if

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \text{then } AB = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

3. Input is a sequence X with n elements that is residing in D disks. The problem is to sort X. It is known that each element in X is an integer in the range $[1, C]$, where C is a constant. Let M be the main memory size. Assume that $M = 2BD$ where B is the block size. Show how to sort X in $O(1)$ (read and write) passes through the data.

4. 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.

5. Input are k polynomials $f_1(x), f_2(x), \ldots, f_k(x)$ with degrees $d_1, d_2, \ldots, d_k$, respectively, with $\sum_{i=1}^{k} d_i = n$. Present an $O(n \log n \log k)$ time algorithm to compute $\Pi_{i=1}^{k} f_i(x)$.

6. Construct a linear regression model for the following input examples: $(0, 1; 6)$, $(1, 0; 2)$, $(1, 1; 5)$, $(1, 2; 10)$. The model of interest is $f(x_1, x_2) = w_1 x_1 + w_2 x_2$. Compute the best values for the parameters w_1 and w_2.