Name: _____

CSE 4502/5717 Big Data Analytics
Exam II; October 31, 2024

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) Input is an undirected graph $G(V, E)$ in the form of adjacency lists. This input is residing in a single disk. The problem is to check if G is a tree. Show how to solve this problem in $O\left(\frac{|V|}{B}\right)$ I/O operations, where B is the block size. You can assume that the core memory M is of size $\Theta(|V|)$.

2. (17 points) Input are $n = 2^k$ (k being an integer) sorted sequences $R_1, R_2, R_3, \ldots, R_n$ each of length M^2, where M is the core memory size. For any i ($1 \leq i \leq n$), the elements in R_i are distinct. These runs are striped across D disks. The problem is to output all the elements that are common to all of the sequences (i.e., $\cap_{i=1}^{n} R_i$). Show how to do this in $k+1$ passes through the data (i.e., $\frac{(k+1)nM^2}{BD}$ parallel I/O operations). Assume that $M = \Theta(BD)$.

3. (16 points) Show that we can sort $M^{4/3}$ keys on the Parallel Disks Model in three passes through the data (assuming that $B = M^{2/3}$). *Hint:* Use the LMM sort algorithm.

4. (16 points) Input are two strings α and β of the same length. The problem is to check if α is a cyclic rotation of β, i.e., α is a suffix of β followed by a prefix of β. For example, if $\beta =$*abcdef*, then $\alpha =$*efabcd* is a cyclic rotation of β. Present an $O(m)$ time algorithm to solve this problem, where $m = |\alpha| = |\beta|$.

5. (17 points) Input is a string S from an alphabet Σ, with $|S| = m$, and an integer k, $1 \leq k \leq m$. Assume that $|\Sigma| = O(1)$ and $k = O(\log m)$. The problem is to output all the distinct k-mers of S, together with a count of how many times each k-mer occurs. For example, if $S = gacagcatgcagatg$ and $k = 3$, then the unique k-mers are $gac, aca, cag, agc, gca, cat, atg, tgc, aga, gat$ and their counts are $1, 1, 2, 1, 2, 1, 2, 1, 1, 1$, respectively. Present an $O(m)$ time algorithm to solve this problem.

6. (17 points) In this problem we are given a text T, a pattern P, and the suffix array S for T. The problem is to identify all the occurrences of P in T. Let $|T| = m$ and $|P| = n$. Present an algorithm to solve this problem in $O(1)$ time using $n\sqrt{m}$ CRCW PRAM processors. Specifically, the output should be an array $A[1 : m]$ such that $A[i] = 1$ if $P = T_i$; (If $T = t_1 t_2 \cdots t_m$ then $T_i = t_i t_{i+1} \cdots t_{i+n-1}$); Also, $A[i] = 0$ if $P \neq T_i$, for $1 \leq i \leq (m - n + 1)$.