Name: _____

CSE 4502/5717 Big Data Analytics
Exam II; November 10, 2022

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 has a triangle. Three nodes a, b, and c in G form a triangle if $(a, b) \in E$, $(b, c) \in E$, and $(c, a) \in E$. Show how to solve this problem in $O\left(\frac{|V|^3}{B}\right)$ 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 4 sorted sequences R_1, R_2, R_3, and R_4 each of length M^2, where M is the core memory size. These runs are striped across D disks. Show how to merge them in two passes through the data. Assume that $M = \Theta(BD)$.

3. (16 points) Show that we can sort $M^{1.25}$ keys on the Parallel Disks Model in three passes through the data (assuming that $B = M^{0.75}$). *Hint:* Use the LMM sort algorithm.

4. (16 points) Input are $(k+1)$ strings $S_1, S_2, \ldots, S_k$ and T, each of length n. The problem is to find the longest substring that is common to $S_1, S_2, \ldots, S_k$ that does not occur in T. Show how to solve this problem in $O(k^2 n)$ time.

5. (17 points) Input are k strings $S_1, S_2, \ldots, S_k$ (with $\sum_{i=1}^{k} |S_i| = M$) and an integer ℓ. The problem is to find a *unique substring*. A string u will be a unique substring if $|u| = \ell$ and u occurs as a substring of exactly one of the k strings. Present an $O(M)$ time algorithm for solving this problem.

6. (17 points) Input is a string $S = t_1 t_2 \cdots t_m$ from an alphabet Σ, where Σ is the set of all integers in the range $[1, m^{10}]$. It is given that the characters in S are distinct. Show how to create a suffix array for S in $O(m)$ time. You cannot use any of the algorithms discussed or cited in class for the construction of suffix arrays.