Data Management – exam of 12/07/2023 **(A)**

Problem 1
The *separation degree* between researcher A and researcher B (different from A) is 1 if they have collaborated, and is 2 if both of them have collaborated with a third researcher C. Let `R(A,B)` be a relation (therefore, without duplicates) storing all the tuples $\langle r, s\rangle$ such that researchers r and s have collaborated. Obviously, `R` is symmetric. We know that (i) no researcher has more than 300 collaborations, (ii) associated to `R` there is a clustering B$^+$-tree index on attribute `A` whose cost for retrieving the first tuple of `R` with a certain value for `A` is 4 page accesses, (iii) every page has space for 500 tuples of `R`, and (iv) the buffer has 100 frames available. Considering the query Q that, given a specific researcher x, computes the set (therefore, without duplicates) of researchers with whom x has separation degree less than or equal to 2, illustrate the most efficient algorithm you can think of for answering Q, and tell which is the cost of the algorithm in terms of number of page accesses.

Problem 2
Consider the following schedule S:

$$r_1(X)\ w_4(Z)\ w_2(X)\ r_2(Y)\ w_3(Y)\ w_3(X)\ r_1(Y)\ r_4(Y)\ w_1(Z)$$

and answer (with suitable motivations) the following questions: (i) Is S conflict serializable? (ii) Is S a 2PL schedule (with shared and exclusive locks)? (iii) Can we insert the commit commands in such a way that the resulting schedule is accepted by the timestamp-based scheduler (assuming that the timestamp associated to a transaction coincides with the physical time when the transaction starts)? (iv) In all the three cases above where the answer is no, provide the answer to the following further question: is there any action α in S such that, if we delete α from S, then the answer to the question would switch to yes?

Problem 3
A schedule S is called *parsimonious* if it contains only write actions and no element of the database is used by more than two actions in S. Prove or disprove that every parsimonious schedule is conflict serializable if and only if it is view serializable.

Problem 4
Let `R(A,B,C)` and `S(D,E)` be two relations, each stored in a heap. A tuple t of `R` is said to be a *lower k-match* with `S` if there are at least k tuples of `S` whose value in the attribute `D` is equal to the value of t in the attribute `C`. Let M be the number of buffer frames available, let Q be the query that, given a value for k, computes all the tuples in `R` that are lower k-matches with `S`, and consider the following questions.

4.1 Describe the conditions under which we can process Q using the one-pass, the two-passes, and the block-nested loop method, respectively.

4.2 For each of the methods mentioned above, illustrate the corresponding algorithm and tell which is the cost of such algorithm in terms of the number of pages of `R` and `S`.

Problem 5 (only for students enrolled in an A.Y. before 2021/22 who do **not** do the project)
Parallel algorithms for the implementation of relational operators are based on the idea of *horizontal data partitioning*, by which we can split a relation in various chunks, each one stored at a different node.

5.1 Describe the techniques that can be used for partitioning.

5.2 Among the described techniques, tell which is the most common one and explain why.

5.3 Illustrate the idea for designing a parallel algorithm that, given a table `S`, based on the partitioning technique mentioned for item 2) above, and depending on the number of buffer frames available at the various nodes, computes the relation obtained from `S` by eliminating duplicates. Also, discuss the cost of the algorithm.

5.4 Illustrate the idea for designing a parallel algorithm that, based on the partitioning technique mentioned for item 2) above, and depending on the number of buffer frames available at the various nodes, computes the set difference of two relations given in input, also discussing the cost of the algorithm.