Department of Computer Science,
New Mexico State University

Qualifying Exam for Databases
Fall, 2014

Open book; Open notes. **Your code:** _____
Good luck! **(2hrs, 100 points)**

Answer **Questions 1–4** by using the following relational database schema, where the underlined attribute(s) form(s) the primary key of the corresponding schema.

- `Scholar(sid:int, sName:varchar(64), email:varchar(32), institute:varchar(64))`
- `Paper(pid:int, pTitle:varchar(128))`
- `Writes(sid:int, pid:int)`
  –Foreign key `sid` references `Scholar(sid)`
  –Foreign key `pid` references `Paper(pid)`
- `Cites(citerPid:int, citeePid:int)`
  –Foreign key `citerPid` references `Paper(pid)`
  –Foreign key `citeePid` references `Paper(pid)`
  - E.g., if a paper with id 1 cites a paper with id 2, citerPid is 1 and citeePid is 2.

Express the following queries in Relational Algebra.

1. (20pts) Write ONE relational algebra expression (NOT a set of relational algebra expressions) to answer each of the following queries.

   A. (10pts) For NMSU scholars (i.e., scholars whose institute is NMSU), find all their non-NMSU collaborators, who co-wrote papers with NMSU scholars and do not work at NMSU. For example, Dr. Son Cao Tran co-wrote a paper with Dr. Enrico Pontelli and Dr. Chitta Baral. Dr. Pontelli should not be treated as a non-NMSU collaborator of Dr. Tran because Dr. Pontelli also works at NMSU. However, Dr. Chitta Baral should be treated as a non-NMSU collaborator of Dr. Tran since Dr. Baral does not work at NMSU. Show the names, emails, and institutes of such non-NMSU collaborators. For each non-NMSU scholar, only show his/her name, email, and institute once even if he/she collaborates several times with multiple NMSU scholars.

   B. (10pts) Assume that a paper is influential if it has been cited more than 1000 times. You are asked to find all the influential papers. For each influential paper, show its title and the number of times that it has been cited.

2. (20pts) Write an SQL statement to answer each of the following queries.

   A. (10pts) Find the scholars who have written more than one hundred research papers. Show such scholars' names, emails, and institutes.

   B. (10pts) For a paper with title "A survey on time series classification", find all the papers that cite this paper and that were not written by any author who wrote "A survey on time series classification". Show the citing papers' titles.

3. (5pts) Suppose that the following function dependencies hold on the *Scholar* relation.

   - $sid \rightarrow sName, email, institute$
   - $email \rightarrow sid, sName, institute$

   Is the *Scholar* relation in BCNF? Please justify your answer.

4. (30pts) Besides the schema information, you are also given the following information about the buffer, disk, and the relations:

   - No record spans multiple disk pages.
   - The *integer* type uses 8 bytes and an index entry pointer uses 16 bytes.
   - Each buffer page size is 1K (i.e., 1024) bytes.
   - Each disk page size is 1K (i.e., 1024) bytes.
   - 10 buffer pages are available.
   - The *Paper* relation contains 100,000 tuples.
   - The *Writes* relation contains 1,000,000 tuples.
   - The *Scholar* relation contains 10,000 tuples. It has a clustered B^+ index on its primary key.

   Note: If the provided information is not sufficient to answer the question and you need to make assumptions, please write down these assumptions clearly.

   A. (15pts) What is the maximum height of the B^+-tree for the *Scholar* relation? Please justify your answer.

   B. (15pts) Given a relational algebra expression "$\sigma_{institute='NMSU'}(Scholar \bowtie Writes)$", please draw the relational algebra tree with your evaluation plan to evaluate this query. Let the block nested loop join algorithm be used, please estimate the cost of your evaluation plan by analyzing (1) the approximate number of results generated in each step and (2) the I/O cost of each step.

5. (25pts) Given the log file as shown in Table 1, and assume that (1) the transaction table is empty at the check point and (2) the dirty page table at the check point contains the following content.

Page	recLSN
P1	80
P2	90

LSN	Log
100	begin_checkpoint
105	end_checkpoint
110	update T2 write P3
115	update T1 write P2
120	update T2 write P2
125	update T1 write P3
130	T1 abort
135	CLR: Undo T1 LSN 125 (undoNextLSN=115)
140	T2 commit
145	Update T3 write P2
150	T2 end
155	CRASH, RESTART

Table 1: Log records

Answer the following questions to perform recovery when DBMS restarts:

A. (10pts) After the analysis phase, please show the content of the dirty page table and the transaction table.

B. (5pts) What is the starting LSN that the DBMS starts to check during the REDO phase?

C. (10pts) After finishing the UNDO phase, show the newly added log records and denote clearly the *undoNextLSN* and *ToUndo*.

— End —