CS432: Assignment 1

(Due in class **at the beginning of class** on 28 September 2007)

Name: _____ Cornell NETID: _____

You may discuss the assignments with other students in the course, but you will have to solve and write up your solutions independently. This assignment will account for 10% of your overall grade. Please write down your answer to each question in the space provided. Show the relevant intermediate steps so that we can give you partial credit even if your final answer is incorrect. Good luck!

Part A: Disk organization (20 points total)

Consider a disk with a sector size of 512 bytes, 8000 tracks per surface, 180 sectors per track, 8 double-sided platters, average seek time of 10 ms, and average rotational delay of 5 ms. Assume that a block size of 1024 bytes is chosen and the size of a page is the same as the size of a block. Now suppose that a file containing 750,000 records of 100 bytes each is to be stored on such a disk and that no record is allowed to span more than one block. Assume it takes 0.2 ms to transfer a block.

A.1) How many disk blocks are required to store the entire file? (5 points)

A.2) If pages are stored sequentially on disk, with page 1 on block 1 on track 1 on the first surface, what is the page stored on block 1 of track 1 on the third disk surface? Briefly explain your answer. (5 points)

A.3) What is the time required to read all the records in the file sequentially? Assume that you can read from only one disk head at a time, and that there is no buffer management in effect. (5 points)

A.4) What is the time required to read all the records in the file in some random order? Assume that you can read from only one disk head at a time, and that there is no buffer management in effect. Further, assume that each read incurs the average seek time and rotational delay. (5 points)

Part B. Buffer Manager (15 points total)

Consider a buffer manager that has a buffer pool large enough to hold 3 pages, and consider a file of size 20 pages. Assume that the buffer pool is initially empty.

B.1) Is it possible to write down a sequence of pin and unpin requests on at least 5 distinct pages so that the state of the buffer pool is the same at the end of the sequence, no matter which of the three replacement policies (CLOCK, LRU, MRU) is used? If yes, write down the sequence, otherwise explain why not. (8 points)

B.2) Give a sequence of pin and unpin requests such that the number of disk accesses is different for CLOCK and FIFO. Show the final states of the buffer pool for each policy. (7 points)

Part C. ER Model (25 points total)

A company is in the business of selling subscriptions to entertainment content over the Internet, and needs you to define a database that meets its needs:

- Each **account** is subscribed to one or more subscription **package**s
- Every **package** provides either music or video, but not both
- Each **package** entitles the user to download a limited set of either **song**s or **video**s, depending on its type
- Each **account** has several associated **user**s, having varying levels of access to the subscriptions associated with the account (e.g., parental controls)
- Each **account** has information for one or more **credit card**s on file
- Every **account** has exactly one primary user
- Every **user** is associated with exactly one account
- Every **credit card** is associated with exactly one account, identified to the user by a short descriptive name

C.1) Draw an ER diagram that captures this information. Be sure to indicate all the key and participation constraints, and any assumptions that you make. (15 points)

C.2) Map this ER diagram to the relational model by writing SQL statements to create the relevant relations. Make sure to capture all the constraints that you can. (Point out explicitly any that you cannot) (10 points)

Part D. Keys (5 points in total)

Consider a relation R (A, B, C, D, E). The only information that you know about R is that AB, BC and CD are keys. List all possible keys (some will not be compatible with others). (5 points)

Part E. Relational Algebra and Domain Relational Calculus (35 points in total)

As part of the global war on terror, the government has been monitoring telecommunications and has recorded a massive amount of data, and now requires your expertise with relational algebra to make sense out of it. Consider the following relational database:

human (id, personName, city, country)
known_terrorist (humanId, funding, prevAttacks : integer)
has_phone (humanId, phoneNumber)
email_account (id, emailAddress, humanId)
phone_call (id, callerNumber, calleeNumber, transcript, date)
email_message (id, senderId, recipientId, subject, body, date)

Thanks to helpful cooperation from the telecom industry, we have a fairly complete picture of who has called whom, but due to rampant lawlessness on the Internet we have at best murky knowledge of how Internet aliases relate to living people, so that the *humanId* field in the **email_account** relation frequently assumes the special value `null`. When writing queries, you may abbreviate attribute names in any reasonably unambiguous manner.

For problems **E.1**, **E.2** and **E.3**, write a query in Relational Algebra.

E.1) Find the names of all people who live in the same city and country as a known terrorist they have contacted, either by phone or email. (5 points)

E.2) We suspect some terrorists are using multiple email addresses. Try to resolve this aliasing by finding all pairs of email accounts (identified by their *id*s) that:

1. have sent messages to exactly the same set of recipients, and
2. have sent at least one message to a known terrorist

Hint: Consider computing the set of all pairs (id1, id2) so that any recipient of a message from id1is also a recipient of a message from id2. (10 points)

E.3) The government believes that the more attacks a terrorist has committed, the more likely s/he is to strike again. Among the known terrorists who have committed the most attacks, find the names of those with the most funding. (10 points)

For problems **E.4** and **E.5**, write a query in Domain Relational Calculus.

E.4) Find the names of all known terrorists who have greater funding than any other known terrorist they have contacted, either by phone or email. (5 points)

E.5) A terrorist *cell* is a group of all the people who have received an email from a common known terrorist. Find all individuals who have phoned every member of some cell. (5 points)

This page will be used for grading your assignment. Do not write on this page.

SECTION	QUESTION	SCORE	SECTION TOTAL
Part A Disk organization	**A.1** (max: 5 points)		(max: 20 points)
	A.2 (max: 5 points)		
	A.3 (max: 5 points)		
	A.4 (max: 5 points)		
Part B Buffer manager	**B.1** (max: 8 points)		(max: 15 points)
	B.2 (max: 7 points)		
Part C ER model	**C.1** (max: 15 points)		(max: 25 points)
	C.2 (max: 10 points)		
Part D Keys	(max: 5 points)		(max: 5 points)
Part E Relational Algebra & Domain Relational Calculus	**E.1** (max: 5 points)		(max: 35 points)
	E.2 (max: 10 points)		
	E.3 (max: 10 points)		
	E.4 (max: 5 points)		
	E.5 (max: 5 points)		
			(max: 100 points)