University of Cape Town
Department of Computer Science

Computer Science CSC303S

November Examination, November 2005

- Please show all your work in arriving at an answer since the reasoning is more important than merely a correct answer.
- Please write the numbers of the questions you answered on the front cover.

Marks: 100

Time: 3 hours

- Approximate marks per question are shown in brackets
- The use of calculators is permitted

Section 1. Information Management

Answer BOTH Questions 1 and 2 and THREE questions from among Question 3 - 6 for a total of 50 marks

Answer this section in a separate book.

Question 1. This question is compulsory [10 marks]

a) Design an Entity-Relationship (ER) model for an estate agency. They wish to keep information on all the properties (e.g. houses) they sell (address, asking price, selling price), on all their customers (name, ID number, telephone number) and on all their agents who sell these properties (name, ID number, telephone number, salary). They also want to keep data on all their sales, so they can have a record of which agent sold which seller's property to which buyer, as well as the date of that sale and the name of the lawyer who did the transaction. A sale of a property involves exactly one buyer and exactly one seller; it can involve one or two agents. Sellers are always customers; sometimes buyers are also customers - if not, they are added as new customers of this agency. A customer can buy/own several properties, and some customers do not own any property at all (yet). The agency will often sell the same property over and over again. Using the ER diagram conventions of your textbook, draw an ER diagram that captures as much of this information as possible. [4]

b) Consider the relation R(A, B, C, D, E, G) for which the following four functional dependencies hold:

```
BD -> G      AB -> CD     AC -> B      AC -> E
```

Convert relation R into an equivalent BCNF (Boyce-Codd Normal Form) schema. Show all your working. [2]

c) Consider the relation scheme S below and then give SQL statements for each of the queries that follow.

```
GAME (GameID, City, Winner, Loser, Points, Revenue )

TEAM (TeamID, Name, Country, Region, HomeCity)
```

i) Find the names of all teams that have lost a game played in the city of London.

ii) Find the total revenue for each city that has hosted more than 2 games.

[4]

Question 2. This question is compulsory [10 marks]

Consider the following collection of documents:

DocumentID	Terms
Doc1	one two three
Doc2	one one two two
Doc3	one
Doc4	three three three three

a) Build inverted files for this document collection. Include a term occurrence count with each DocumentID listed. [3]

b) If the query "two three" is submitted to a Boolean-OR-based search engine, which results will remain after filtering? [1]

c) Using the following ranking formula, compute a rank for each result from the previous question.

$$\text{Similarity} = \frac{1}{|D|} \sum_{t \in Q \cap D} (1 + \log_e f_{d,t}) \cdot \log_e (1 + \frac{N}{f_t})$$

Assume that:

- D is the length of the document, including all terms in the document.
- Only terms common to both query and document are considered.
- N = the total number of documents in the result set.
- $f_{d,t}$ = term frequency of term t in document d.
- f_t = number of documents term t appears in.

[3]

d) Discuss how the inverted files can be optimised to use less space if the number of documents (and therefore document identifiers) is large. [2]

e) Briefly discuss one technique that can be used to improve on recall. [1]

Answer THREE Questions from 3, 4, 5 and 6

Question 3. [10 marks]

a) For the relation scheme S in ¡question 1e¿ above, give relational algebra expression(s) for each of the queries below:

i) Find the names of all teams that have lost a game played in the city of London.

ii) Find out which teams have won every ”home” game they played (i.e. have won every game they played where the City played in was in fact their own HomeCity).

iii) Find all the Revenue values associated with the games in which the ”Bulls” team played - i.e. either as Winner or as Loser.

[5]

b) Briefly explain the difference between the two terms in each of the following pairs:

i) logical data independence and physical data independence

ii) a procedural query language and a non-procedural query language

iii) atomicity and concurrency

iv) sophisticated users and specialised users (of database systems)

v) the transaction manager and the query processor (in a database management system)

[5]

Question 4. [10 marks]

a) For the relation scheme S in Question 1c above, give SQL statements for each of the queries below:

i) Find the name of the team(s) which scored the most points (in total, over all games).

ii) Find the total Revenue collected in each country. Note that every City in which a game is played is the HomeCity of some team.

iii) Find the names of those teams that have won more points (in total, over all games) than they have conceded (in total, over all games). The number of points conceded in a game is the number of points awarded in a game where that team was the Loser.

[5]

b) Consider the relation R(A, B, C, D, E, G, H) for which the following functional dependencies hold:

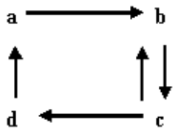
```
BD -> G      AB -> C     D -> B      AD -> CE   C -> B
```

i) Show any two changes you would make to the given set of dependencies if you were asked to find a canonical cover. As it turns out, only two changes are needed, so your result is in fact a canonical cover. [2]

ii) Use your canonical cover from part (i) to find a 3NF (3rd Normal Form) schema for the seven attributes of relation R. (Note, if you omitted the first part of this question, use the given set of dependencies as your canonical cover and continue from there). Show all your working. [3]

Question 5. [10 marks]

a) Apply the PageRank algorithm to the following link graph.



Start with equal ranks of 1/4 each. Stop after 4 iterations, or when C=5/16. Show all calculations, including N(umber of forward links) values and B(ack links) sets. (Hint: Remember that

$$r[i]_n = \sum_{j \in B[i]} \frac{r[j]_{n-1}}{N[j]}$$

[6]

b) How does the HITS algorithm differ conceptually from the PageRank algorithm? [2]

c) Contrast the runtime performance of simple PageRank with simple HITS. [2]

Question 6. [10 marks]

a) Write an XSLT template to transform

```
<people>
<name age="21">wert</name>
<name age="23">polk</name>
<name age="22">jik</name>
</people>
```

into

```
<table>
<tr><th>Names</th><th>Ages</th></tr>
<tr><td>wert</td><td>21</td></tr>
<tr><td>polk</td><td>23</td></tr>
<tr><td>jik</td><td>22</td></tr>
</table>
```

Assume that the values such as ”wert” may differ from one document to another. Assume the source namespace prefix is ”source” and the destination prefix is ”dest”. Assume that the number of ¡name¿ tags is unbounded in its defining XML Schema. Use the following as a starting point.

```
<xslt:template match="source:people">
. . .
</xslt:template>
```

[5]

b) Write an XML Schema type definition corresponding to the contents of the people node. [5]

Section 2. Theory of Algorithms

Answer BOTH Questions 7 and 8 and THREE out of the last 4 questions for a total of 50 marks

Answer this section in a separate book.

Question 7. This question is compulsory [10 marks]

a) i) Briefly explain the term ”dynamic programming”.

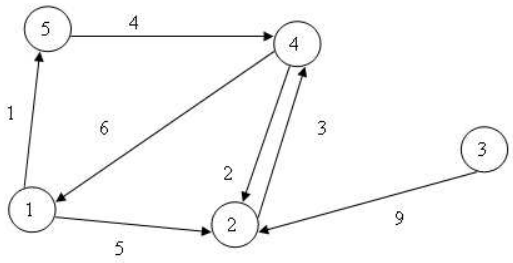
ii) This technique can be improved by using ”memory functions”. What weakness of dynamic programming does this address, and how

[3]

b) Given the graph below, show how

- **either** Floyd’s algorithm for all-pairs shortest-paths
- **or** Warshall’s algorithm for transitive closure

would be applied to it. State which algorithm you are applying, and then show all your working.



[4]

c) In the minimum-coins-change problem, an amount of change N must be given using as few coins as possible, and coins have M different values/denominations v1, v2, ., vM (e.g. for our money, M is 9 and the 9 (cent) values are: 1, 2, 5, 10, 20, 50, 100, 200, 500). Consider how you would use dynamic programming to solve this problem and then answer the following questions:

i) In one or two sentences, outline how you would apply dynamic programming to solve this problem (brief top-level description).

ii) Apart from an array to store the M denominations (e.g. 1, 2, 5, ., 500), you are probably using another array as well in your solution. If you are not using any other array, explain why not. If you are using another array, give a pseudocode statement to show how you would initialize an element of this array at the beginning of the algorithm, and give a comment to explain (what is the meaning of this array element, and why must it start off at that initial value).

[3]

Question 8. This question is compulsory [10 marks]

a) Draw a diagram to illustrate the process of solving algorithmic problems. Describe what each step in your diagram entails. [6]

b) Binary search is often incorrectly classified as a divide-and-conquer algorithm. Why is this incorrect and what is the correct classification? [2]

c) What is amortized efficiency and why is it an important form of analysis? [2]

Answer THREE Questions from 9, 10, 11, 12 and 13

Question 9. [10 marks]

a) i) Explain what it means when a problem has a ”tight lower bound”.

ii) Does an NP-complete problem have a tight lower bound (Yes/No)? Give a brief reason for your answer.

[3]

b) Consider the following problems, which are all known to be NP-complete:

- **[HC] Hamiltonian Circuit:** Does a given graph have a path that starts and ends at the same vertex and passes through all the other vertices exactly once)?
- **[TS] Traveling Salesman:** Is there a path through a given weighted graph that starts and ends at the same vertex and passes through all the other vertices exactly once, and has a total weight less than W?
- **[DP] Disjoint paths:** For a given directed graph, is there a pair of disjoint (non-overlapping) simple paths through the graph such that the one path starts at vertex A and ends at vertex B, and the other starts at vertex C and ends at vertex D? (a simple path is one that passes through each vertex exactly once).
- **[DE] Distinguished Edge:** For a given directed graph, is there a simple path from vertex A to vertex B that passes along a specific edge E

Show that

- **either** TS (Traveling Salesman)
- **or** DE (Distinguished Edge)

is NP-complete. In your answer, you may use the fact that any of the other three problems above are known to be NP-complete

[5]

c) Let X be a new problem. Suppose a computer scientist can show that X is an NP (nondeterministic polynomial) problem and can also prove that its lower bound is exponential. Which two of the following will then be true? Give a reason for each answer.

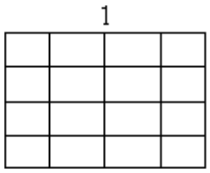
- X is NP-Complete
- X is intractable
- X is undecideable
- HC (Hamiltonian Circuit problem) is not in P
- NP = P

[2]

Question 10. [10 marks]

a) Explain the backtracking and the branch-and-bound techniques of algorithm design briefly but in such a way that the distinction between them is clear. [2]

b) Suppose we use backtracking to solve the 4-Queens problem. This problem is: find any one way of placing 4 Queens on a 4-by-4 chess board (the board is shown below) in such a way that no two Queens are in the same column or in the same row or on the same diagonal. The 4-by-4 board below would be at the root of the state-space-tree. Apply the algorithm for just a few steps, stopping once you have drawn the next 3 (three) boards (excluding the root) that would be generated using backtracking. Indicate with an X wherever you detect an unpromising state. Label each node of your tree (label the X nodes that are non-promising as well as the promising nodes which have a board associated with them) to indicate the order in which you have generated your tree. For example, the root node below has been labeled with a 1 as it is the first node generated; now proceed from there..



[3]

c) The IsEquals(x,y) function for natural numbers x and y, which returns 1 if x == y and returns 0 otherwise, is clearly computable.

i) **EITHER**design a Turing Machine for IsEquals

ii) **OR** show that IsEquals is a Partial Recursive function

If you do (i) the Turing Machine, use the unary system i.e. let a natural number N be represented as a sequence of (N+1) 1's Assume that the input is given with an asterisk (*) in the leftmost cell, followed by the first number x, followed by an asterisk (*), followed by the second number y, followed by a hash (#) to indicate the end of the input. Your Turing Machine must leave its answer (0 or 1) on the cell immediately to the right of the hash symbol. Example:

```
Input tape:         *1111*111#

Output tape:       *1111*111#0    (no, the two numbers are unequal)
```

Note: the cells to the left of the hash can be left as they are or can be changed to any other symbol, so e.g. another valid output for the above input would be:

```
Output tape:       *$$$$*$$$#0    (no, the two numbers are unequal)
```

If you do (ii) the Partial Recursive function proof, you may assume that pred (predecessor) is a Partial Recursive function

```
where Pred(x) = x-1 for x > 0  and Pred(0) = 0
```

[5]

Question 11. [10 marks]

Consider the following algorithm:

```
Algorithm: Recurse(A[l..r])
// Input: An array A[l..r] of real numbers
if l = r return A[l]
else
    temp1   Recurse(A[l..  (l+r)/2  ])
    temp2   Recurse(A[  (l+r)/2  +1 .. r])
    if temp1 <= temp2
        return temp1
    else
        return temp2
```

a) What is this algorithm's basic operation? [1]

b) How do the worst-, average- and best-case efficiencies of this algorithm differ? [1]

c) Find the time efficiency of the algorithm by setting up a recurrence relation and solving it by backward substitution. [7]

d) What does this algorithm compute? [1]

Question 12. [10 marks]

a) What are the relative strengths and weaknesses of the theoretical and empirical approaches to analysing the time efficiency of an algorithm? [4]

b) Given a polynomial (with degree n) of the form $p(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$

i) Provide two algorithms for evaluating this polynomial for a given x = c. [4]

ii) What are the time efficiencies of your two algorithms and which is better? [2]

Question 13. [10 marks]

You are given an alphabet and associated probabilities of occurrence, as follows:

Character	A	B	C	D	E	_
Probability	0.35	0.2	0.15	0.05	0.2	0.05

a) Construct a Huffman Coding Tree and Huffman Codes for this data. Be sure to show all the intermediate steps of Huffman's algorithm [10]