ENGINEERING TRIPOS PART IIA

Monday 30 April 2007 2.30 to 4

Module 3I1

DATA STRUCTURES AND ALGORITHMS

*Answer **all** of Section A (which consists of short questions), and **two** questions from Section B.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

STATIONERY REQUIREMENTS	SPECIAL REQUIREMENTS
Single-sided script paper	none

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

SECTION A

Answer all parts of this question. The question in this section will be marked out of the same total as each question in section B, and each part carries the same weight.

1 (a) A graph has all its edges of length 1. Explain how you would find a minimum-cost spanning tree for it, commenting on any ways you exploit your knowledge about its edge-lengths. [10%]

(b) Why do we normally say that sorting is an $n\log(n)$ process? [10%]

(c) A Huffman-style coding will be used to process N symbols each of which consists of 8-bits. In other words it will work over an alphabet of size 256. Rather than counting character frequencies in the material to be handled a fixed coding tree will be used. What is the greatest compression that could possibly be obtained, and what is the worst expansion that could be experienced? In each case explain briefly when and how the case arises. [10%]

(d) What does it mean for a sorting method to be *stable*? Which, if any, of Quicksort, Shell's Sort and Heapsort is liable to be stable if implemented simply? [10%]

(e) For a buddy system of free-store administration explain the steps that are taken when a block of store is to be released for re-cycling. [10%]

(f) Explain the terms *directed graph*, *undirected graph* and *bipartite graph*. [10%]

(g) Explain how to add a single new item to a *heap*, preserving the heap property. What is the cost of this operation? [10%]

(h) You are going to compress a document using Lempel Zif (LZ) compression. The document is a slightly updated version of one you compressed and sent to the recipient last week. Explain why it could make sense to run the previous version through the compressor, discarding the compressed output, before processing the new version. How would the recipient have to act to recover the text? [10%]

(cont.

(i) Explain one benefit and one limitation of a garbage collection scheme that works by copying all active data. [10%]

(j) What is a b-tree and how would you look up information in one? You do not need to explain how to create or modify a b-tree. [10%]

SECTION B

*Attempt **two** questions from this section. Each question has the same weight in marks.*

2 (a) Suppose that a file-sharing service on the internet has several million subscribers. Some internet service providers attempt to block the network, and so at any one moment only certain pairs of users can exchange information directly. The file sharing software will try to pass on material by fetching it indirectly via other users when that is possible. Explain an algorithm that (supposing you know exactly which pairs of nodes can communicate) can find the shortest number of hops to get data from user A to user B. Explain how its costs will scale as the network grows. [30%]

(b) An upgrade to the file-sharing software now wants to route information to maximise speed. If assigns each point-to-point link a "cost" based on how long it takes a transfer a file between them. If an intermediate node does not start forwarding data until it has finished receiving it the cost associated with a sequence of links is just the sum of the individual link costs. Explain how to find the cheapest route from A to B in this model. Again give cost predictions for the calculation and discuss how it might scale with network size. [40%]

(c) For a rather smaller network you have been challenged to produce a complete table showing the expected costs of communication between all pairs of nodes. But this time your cost model will suppose that intermediate nodes can start retransmitting material as soon as they receive anything, so for instance the cost associated with a path A-B-C-D will now be the greatest of the individual costs A-B, B-C and C-D. Invent an algorithm that will find all the costs in this case, and analyse its complexity. [30%]

3 (a) "treesort" works by adding the items that are to be sorted to an ordered binary tree and then flattening that tree. Explain why the numbers of comparisons performed will be the same as the number done by a variant of quicksort. [25%]

(b) What is a splay tree? What are their properties and why might they be preferred to some other related data structures? You are not required to give full details of the exact transformations that maintain splay trees. [25%]

(c) A splay tree is made by starting with an empty tree and adding items in ascending order. No look-up operations are performed, just the ones that add n items in order. Explain carefully what happens as each item is added. Show what shape the tree ends up and discuss the total cost of building it. [25%]

(d) "splaysort" works by taking a sequence of items that are to be sorted and adding them one at a time into an initially empty splay tree. A simple traversal of that tree (not changing it at all while traversing it) can then read off the items left-to-right in sorted order on $O(n)$ time. Make predictions about the cost and practicality of splaysort as compared with the use of quicksort. [25%]

(TURN OVER

4 (a) A computer has the text of this examination question and it is going to try to find the word "examination" in it. It will so the search by first checking the final letter of the word searched for. So its initial step will be to compare the underlined characters as here:

•`A computer_has ...`

•`examination`

Working by hand show how the search can proceed from there and comment on how well looking at the final character first helps speed the search up. You do not need to explain in detail how to set up any tables that your method would use. [25%]

(b) Instead of comparing individual characters it will now try a method based on a hash value for the target word. Explain how this can work and discuss how the cost will depend on the lengths of both the word sought and the amount of text that it is to be looked for within. [25%]

(c) A sequence of k independent hash-values are computed, each in the range from 0 to $N-1$. What is the probability that they are all different? Given that when j is much smaller then N, $(1-j/N)$ has about the same value as $(1-1/N)^j$ and that $(1-1/N)^N$ is close to $1/e = 1/2.718 = 0.367$ give a rough estimate, in terms of N, of how large k should be to give a reasonable chance that there has been a hash-collision. [25%]

(d) A programmer generates a sequence that is expected to be "random" by starting with a seed value r_0. At each stage they obtain the next number in their sequence by computing a hash value on the current one. Their way of finding a hash function that works on integers is to start by converting the integer to a string, as if for printing, and then hashing that string. All their integers are kept as 32-bit values. Comment on this scheme. [25%]

END OF PAPER