COMPUTER SCIENCE TRIPOS Part II

Thursday 6 June 2013 1.30 to 4.30

COMPUTER SCIENCE Paper 9

Answer ***five*** *questions.*

Submit the answers in five ***separate*** *bundles, each with its own cover sheet. On each cover sheet, write the numbers of* ***all*** *attempted questions, and circle the number of the question attached.*

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

STATIONERY REQUIREMENTS	SPECIAL REQUIREMENTS
Script paper	*Approved calculator permitted*
Blue cover sheets	
Tags	

1 Bioinformatics

(*a*) What are the usage and the limitations of the Bootstrap technique in phylogeny? [6 marks]

(*b*) We often use Hidden Markov Models (HMM) to predict a pattern (for instance the exons). How can you compute the number of True Positives, True Negatives, False Positives and False Negatives and use them to evaluate your HMM? [6 marks]

(*c*) How can you evaluate the results obtained (number of clusters and their relative position) using the K means algorithm for clustering? [5 marks]

(*d*) What is the difference between the adjacency list and the accessibility list? [3 marks]

2 Computer Systems Modelling

(*a*) Define a Poisson process of rate $\lambda > 0$. [3 marks]

(*b*) Show that the number of events, $N(t)$, of a Poisson process that occur in the fixed time interval $[0, t]$ is a random variable that has a Poisson distribution with parameter λt. [3 marks]

(*c*) Show that the inter-event times of a Poisson process form a sequence of independent random variables each distributed with an exponential distribution with parameter λ. [3 marks]

(*d*) Describe how to use the inverse transform method to simulate exponential random variables with parameter λ. [3 marks]

(*e*) Show how your simulated exponential random variables can be used to simulate Poisson random variables with parameter λ. [3 marks]

(*f*) Consider positive numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ and weight factors $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $\alpha_i \geq 0$ for $i = 1, 2, \ldots, n$ and $\sum_{i=1}^{n} \alpha_i = 1$. Show that

$$f(x) = \begin{cases} \sum_{i=1}^{n} \alpha_i \lambda_i e^{-\lambda_i x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

is a density for a random variable and describe a procedure to simulate values from this density. [5 marks]

3 Computer Vision

(*a*) Explain why inferring object surface properties from image properties is an ill-posed problem in general. In the case of inferring the colours of objects from images of the objects, how does knowledge of the properties of the illuminant affect the status of the problem and its solubility? [4 marks]

(*b*) What surface properties can cause a human face to form either a Lambertian image or a specular image, or an image lying anywhere on a continuum between those two extremes? In terms of geometry and angles, what defines these two extremes of image formation? What difficulties do these factors create for efforts to extract facial structure from facial images using "shape-from-shading" inference techniques? [4 marks]

(*c*) Explain and illustrate the "Paradox of Cognitive Penetrance" as it relates to computer vision algorithms that we know how to construct, compared with the algorithms underlying human visual competence. Discuss how human visual illusions may relate to this paradox. Comment on the significance of this paradox for computer vision research. [4 marks]

(*d*) Define the "Correspondence Problem", detailing the different forms that it takes in stereo vision and in motion vision, and discuss its complexity. In both cases, explain why the computation is necessary. What are the roles of space and time in the two cases, and what symmetries exist between the stereo and motion vision versions of the Correspondence Problem? [4 marks]

(*e*) When defining and selecting which features to extract in a pattern classification system, what is the goal for the statistical clustering behaviour of the data in terms of the variances within and amongst the different classes? [4 marks]

4 Denotational Semantics

(*a*) (*i*) State carefully, without proof, the compositionality, soundness, and adequacy results for PCF. [6 marks]

(*ii*) Define the notion of contextual equivalence in PCF. [2 marks]

(You need not describe the syntax and the operational and denotational semantics of PCF.)

(*b*) Show that for all types τ and closed terms M and M' of type τ, if $[\![M]\!]$ and $[\![M']\!]$ are equal elements of the domain $[\![\tau]\!]$ then M and M' are contextually equivalent. [4 marks]

(*c*) Consider the following closed PCF terms of type $nat \to bool \to nat$:

$$\begin{array}{lcl} F_0 & = & \mathbf{fn}\ x : nat.\ \mathbf{fn}\ y : bool.\ x \\ F_1 & = & \mathbf{fix}\big(\ \mathbf{fn}\ f : nat \to bool \to nat.\ \mathbf{fn}\ x : nat.\ \mathbf{fn}\ y : bool. \\ & & \qquad\quad \mathbf{if}\ \mathbf{zero}(x)\ \mathbf{then}\ \mathbf{0} \\ & & \qquad\quad \mathbf{else}\ \mathbf{succ}(\ f\,(\mathbf{pred}\ x)\,y\ )\ \big) \\ F_2 & = & \mathbf{fn}\ x : nat.\ \mathbf{fn}\ y : bool.\ \mathbf{if}\ y\ \mathbf{then}\ x\ \mathbf{else}\ x \end{array}$$

State whether or not F_1 and F_2 are contextually equivalent to F_0. Justify your answers. [4 marks each]

5 Digital Signal Processing

(*a*) Consider a digital filter with impulse response

$$h_i = 2\alpha \cdot \frac{\sin[2\pi(i - n/2)\alpha]}{2\pi(i - n/2)\alpha} \cdot w_i \quad \text{where} \quad w_i = \begin{cases} 1, & 0 \le i \le n \\ 0, & \text{otherwise} \end{cases} .$$

(*i*) What type of filter is this? [4 marks]

(*ii*) How are the sampling rate f_s at which this filter is operated and its −6 dB cut-off frequency f_c related to parameter α? [2 marks]

(*b*) In an open-source audio-effect library, you find a C routine for processing a recorded voice to sound like it came over an analog phone line:

```
#include <math.h>
#define N 512
#define PI 3.14159265358979323846
void phone_effect(double *x, double *y, int m)
{
  double w, p, f, g, h[N+1];
  int i, k;
  for (i = 0; i <= N; i++) {
    w = 0.54 - 0.46 * cos(2*PI*i/N);
    p = 2 * PI * (i-N/2) / 10;
    f = w * ((p == 0) ? 1 : sin(p)/p) / 5;
    p = 2 * PI * (i-N/2) / 100;
    g = w * ((p == 0) ? 1 : sin(p)/p) / 50;
    h[i] = f - g;
  }
  for (i = 0; i < m; i++) {
    y[i] = 0;
    for (k = 0; k <= N && k <= i; k++)
      y[i] += x[i - k] * h[k];
  }
}
```

The input array x and the output array y each hold m samples of an audio recording (mono) at sampling frequency $f_\mathrm{s} = 32$ kHz.

(*i*) Explain in detail what operation is implemented here (e.g., type of filter, order, cut-off frequency) and how it has been constructed. [8 marks]

(*ii*) You want to use this algorithm on audio recordings with a sampling rate of 48 kHz. What do you have to change in the source code to ensure that the audible effect remains the same? [6 marks]

6 Information Theory and Coding

(*a*) Two random variables X and Y are correlated. The marginal probabilities $p(X)$ and $p(Y)$ are known, as is their joint probability $p(X,Y)$. Give an expression for the conditional probability $p(X|Y)$ using the known quantities. Then, using $p(X)$, $p(Y)$, and $p(X|Y)$, give an expression for the information gained, in bits, from observing Y after X was already observed. [2 marks]

(*b*) Let the random variable X be five possible symbols $\{\alpha, \beta, \gamma, \delta, \epsilon\}$. Consider two probability distributions $p(x)$ and $q(x)$ over these symbols, and two possible coding schemes $C_1(x)$ and $C_2(x)$ for this random variable:

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
α	1/2	1/2	0	0
β	1/4	1/8	10	100
γ	1/8	1/8	110	101
δ	1/16	1/8	1110	110
ϵ	1/16	1/8	1111	111

(*i*) Calculate $H(p)$, $H(q)$, and relative entropies (Kullback-Leibler distances) $D(p||q)$ and $D(q||p)$. [4 marks]

(*ii*) Show that the average codeword length of C_1 under p is equal to $H(p)$, and thus C_1 is optimal for p. Show that C_2 is optimal for q. [2 marks]

(*iii*) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords? By how much does it exceed the entropy $H(p)$? Relate your answer to $D(p||q)$. [2 marks]

(*iv*) If we use code C_1 when the distribution is q, by how much does the average codeword length exceed $H(q)$? Relate your answer to $D(q||p)$. [2 marks]

(*c*) Compare and contrast the compression strategies deployed in the JPEG and JPEG-2000 protocols. Include these topics: the underlying transforms used; their computational efficiency and ease of implementation; artefacts introduced in lossy mode; typical compression factors; and their relative performance when used to achieve severe compression rates. [5 marks]

(*d*) Discuss the following concepts in Kolmogorov's theory of pattern complexity: how writing a program that generates a pattern is a way of compressing it, and executing such a program decompresses it; fractals; patterns that are their own shortest possible description; and Kolmogorov incompressibility. [3 marks]

7 Mobile and Sensor Systems

Duke of Cambridge College has invested in a number of sensors for a pilot study to monitor the use of its spaces (both outdoor and indoor), its energy efficiency and the comfort of its students. In particular, the college has invested in temperature sensors, light sensors and movement sensors, all equipped with WiFi and Zigbee radio chips (assume the college has a WiFi network in place already). The college council has, however, prohibited the installation of permanent wires for powering the sensors in the interest of period building preservation.

(*a*) Describe the architecture for the wireless sensor network system you would put in place on behalf of the college, illustrating the physical components and the communication infrastructure. [3 marks]

(*b*) Select a wireless sensor network MAC protocol for this network and explain why you have chosen it. [3 marks]

(*c*) Describe the *Directed Diffusion* protocol and illustrate how the protocol can be applied in the wireless sensor network of Duke of Cambridge College. [5 marks]

(*d*) Now assume that due to connectivity limitations and the extension of the college grounds, a group of nodes of the wireless sensor network cannot be connected to the rest of the network, nor the college Internet. The data generated by these nodes, however, is vital but not time critical. Explain how you would handle the data harvesting for this portion of the network through the use of a delay tolerant networking approach. [4 marks]

(*e*) After the first roll-out of the pilot study, a bright student suggests that a mobile phone application could be released to allow willing students who want to install it on their smartphones to contribute data about their use of college space. Explain what salient features the application would need to have and how it would be able to detect the student's use of space (mainly location and activity) and transmit the data back to the college system. [5 marks]

8 Natural Language Processing

Doc 1:
... a large variety of plant life or flora in South Africa...
... an entire plant kingdom inside its borders ...
... more plant species than the entire of the UK...
... related sets of plant and animal life...
... the largest genus of flowering plants...

Doc 2:
...tomato processing plant ...
...plant technicians jobs at P&P...
...process operations of the plant...
...plant management...
...CEO of the plant...

Doc 3:
...plants' ability to adapt to harsh circumstances...
...half human, half plant...
...traditional food plants...
...tomato plant...
...spectacular plant...
...medicinal plants...

Doc 4:
...popular magazine "Plant Life", which humourously describes daily life at the Imperial food processing plant...
...published by a chemical plant process worker...
...small molecular commercial plants...
...pharmaceutical plants...
...heavy plant crossing...

(*a*) How many senses of "plant" can you identify in the snippets from the 4 documents given above? Describe each sense by giving a hypernym. [2 marks]

(*b*) Describe how Yarowsky's algorithm for word sense disambiguation would process the example texts. Illustrate each stage of the algorithm with an example. [6 marks]

(*c*) Under adverse conditions, the algorithm in part (*b*) can rapidly diverge from a good solution. What are these conditions? Illustrate using the examples above. [4 marks]

(*d*) How could you apply the Naive Bayes machine learning algorithm, discussed in the lectures in combination with several tasks, as a method for learning the word senses for an ambiguous word? Give the formula and explain how the necessary parameters can be trained. [8 marks]

9 Optimising Compilers

(*a*) "Register allocation and transformation to SSA form are inverses as both use live ranges." Discuss this statement. [5 marks]

(*b*) Explain the problems which can arise in classical dataflow analysis when some variables may hold the addresses of other variables. Indicate the connection, if any, of this problem to that of constructing a call graph for a program which contains higher-order functions. [5 marks]

(*c*) There are two ways to formalise points-to analysis: one is a dataflow analysis (which you need not formulate except for indicating whether the analysis is forward or backward) in which the dataflow values ascribed to each program point are sets of pairs (pointer, pointee); the other is Andersen's analysis. Distinguish these analyses, comparing the precision of their results using an example (it suffices merely to state the analysis results for your example rather than explicitly solving equations). When might the less precise analysis be preferred? [5 marks]

(*d*) What are strictness functions and how do they differ in expressiveness from a list of parameters which a function may receive by value? Give, if possible, source language functions which have $\lambda(x, y).x \wedge y$, $\lambda(x, y).x \vee y$, $\lambda(x, y).x$ and $\lambda(x, y).\neg x$ as strictness functions. [5 marks]

10 Principles of Communications

The modern smart phone features voice (and video) and data communication. However, in the first three generations of technology (from the initial cellular systems up until 3G phones), the network for voice has been separate from the one for data. The GSM circuit-switched part uses a customized protocol stack for carrying the time critical speech traffic.

The next generation of networks (*"4G"*) are intended to be integrated around the Internet Protocol. What are the technical enhancements required to a best-effort Internet Service, to support the different types of flows while meeting the users' needs? [20 marks]

11 System-on-Chip Design

(*a*) List a set of parallel wires that might form a parallel, synchronous interface for transferring 4-bit words when both sides are ready to communicate. [3 marks]

(*b*) Describe in words the protocol used over the interface of part (*a*). Use a timing diagram as well if helpful. [3 marks]

(*c*) Give the circuit for a receiver that is conformant to your interface and protocol from part (*a*) and part (*b*). It should simply store the received data in a register. Use Verilog-style RTL instead of a circuit diagram if you wish. [3 marks]

(*d*) Give, with justification, the maximum throughput of your protocol above in terms of words per clock cycle. [1 mark]

(*e*) A FIFO component (first-in, first-out queue) has two instances of the above interface that share a common clock. Can it be designed so that there is no combinational logic path between the two instances? What does this mean for its behaviour? [5 marks]

(*f*) A variation on the FIFO component has independent clocks for the two instances of the interface. It is used for transferring data between two clock domains. How would it be designed internally and what is its maximum throughput? [5 marks]

12 Topical Issues

(*a*) Compare and contrast the Active Bat system with GPS. [6 marks]

(*b*) An alternative ultrasonic positioning system could use network-connected beacons positioned around a building that simultaneously emit a radio signal and an ultrasonic pulse. Describe how this would work and discuss its advantages and disadvantages over the Active Bat system. [8 marks]

(*c*) The Active Bat system makes use of narrowband ultrasound for the propagation medium. It can be adapted to use wideband ultrasound signals. By analogy with the benefits that wideband radio signals give GPS, discuss the advantages and disadvantages of doing so. [6 marks]

13 Topics in Concurrency

This question is on HOPLA and PCCS, a variant of pure CCS in which any output on a channel persists. Let A be a set of channel names ranged over by a, b, c and let $\bar{A}$ be the set of complemented channel names, $\bar{A} = \{\bar{a} \mid a \in A\}$. The set of labels $L = A \cup \bar{A}$ is ranged over by l, to which we extend complementation by taking $\bar{\bar{l}} = l$. Use α to range over $L \cup \{\tau\}$, where τ is a distinct label. The terms of PCCS follow the grammar $P ::= \mathbf{nil} \mid \bar{a} \mid a.P \mid (P_1 \parallel P_2)$. The operational semantics of PCCS is:

$$\frac{}{\bar{a} \xrightarrow{\bar{a}} \bar{a}} \qquad \frac{}{a.P \xrightarrow{a} P} \qquad \frac{P_1 \xrightarrow{\alpha} P_1'}{P_1 \parallel P_2 \xrightarrow{\alpha} P_1' \parallel P_2} \qquad \frac{P_2 \xrightarrow{\alpha} P_2'}{P_1 \parallel P_2 \xrightarrow{\alpha} P_1 \parallel P_2'} \qquad \frac{P_1 \xrightarrow{l} P_1' \quad P_2 \xrightarrow{\bar{l}} P_2'}{P_1 \parallel P_2 \xrightarrow{\tau} P_1' \parallel P_2'}$$

(*a*) Draw the transition system of the PCCS term $\bar{a} \parallel a.a.\bar{b}$. [3 marks]

(*b*) This part of the question is on HOPLA. For reference, the operational semantics of HOPLA is presented at the end of the question.

(*i*) For u of sum type, let $[u > a.x \Rightarrow t]$ abbreviate $[\pi_a(u) > .x \Rightarrow t]$. Derive a rule for the transitions of $[u > a.x \Rightarrow t]$. [2 marks]

(*ii*) Show that $[a.u > a.x \Rightarrow t] \sim t[u/x]$ and $[a.u > b.x \Rightarrow t] \sim nil$ if $a \neq b$, where nil represents the empty sum and $\sim$ is the bisimilarity of HOPLA. [4 marks]

(*c*) Write down a HOPLA term realising the parallel composition of PCCS. Use this to give an encoding of PCCS into HOPLA, specifying a HOPLA term $[\![P]\!]$ for every PCCS term P. [*Hint:* The realisation of parallel composition should be the same as that of the encoding of pure CCS into HOPLA.] [5 marks]

(*d*) Use the rules of HOPLA to show how a derivation establishing $[\![P_1 \parallel P_2]\!] \xrightarrow{\alpha.} [\![P_1' \parallel P_2]\!]$ can be constructed from a derivation of $[\![P_1]\!] \xrightarrow{\alpha.} [\![P_1']\!]$.

Explain briefly how you would show that if $P \xrightarrow{\alpha} P'$ in PCCS then $[\![P]\!] \xrightarrow{\alpha} [\![P']\!]$ in HOPLA. In what part of the proof would the derivation that you have constructed be useful?

[6 marks]

Subject to suitable typings, HOPLA has transitions $t \xrightarrow{p} t'$ between closed terms t, t' and action p given by the following rules:

$$\frac{t[rec\,x\,t/x] \xrightarrow{p} t'}{rec\,x\,t \xrightarrow{p} t'} \qquad \frac{t_j \xrightarrow{p} t'}{\sum_{i \in I} t_i \xrightarrow{p} t'}(j \in I) \qquad \frac{}{.t \xrightarrow{.} t} \qquad \frac{u \xrightarrow{.} u' \quad t[u'/x] \xrightarrow{p} t'}{[u > .x \Rightarrow t] \xrightarrow{p} t'}$$

$$\frac{t[u/x] \xrightarrow{p} t'}{\lambda x\,t \xrightarrow{u \mapsto p} t'} \qquad \frac{t \xrightarrow{u \mapsto p} t'}{t\ u \xrightarrow{p} t'} \qquad \frac{t \xrightarrow{p} t'}{a\,t \xrightarrow{a\,p} t'} \qquad \frac{t \xrightarrow{a\,p} t'}{\pi_a(t) \xrightarrow{p} t'}$$

14 Types

(*a*) Give the Mini-ML typing rule for expressions of the form $\texttt{let}\, x = M_1 \,\texttt{in}\, M_2$. How and why is this rule modified in the full ML language? [5 marks]

(*b*) Given a Mini-ML typing problem $\Gamma \vdash M : ?$, define what is a solution for it and what it means for a solution to be principal. [3 marks]

Do the following Mini-ML typing problems have solutions? Justify your answer in each case.

(*i*) $f : \forall \{\} \, (\alpha \to \beta) \vdash (f \, \texttt{true}) \, f : ?$ [3 marks]

(*ii*) $f : \forall \{\beta\} \, (\alpha \to \beta) \vdash (f \, \texttt{true}) \, f : ?$ [4 marks]

(*iii*) $f : \forall \{\} \, (\alpha \to \beta) \vdash \texttt{let}\, f = \lambda x(\lambda y(y)) \,\texttt{in}\, (f \, \texttt{true}) \, f : ?$ [5 marks]

END OF PAPER