COMPUTER SCIENCE TRIPOS Part IA 75%, Part IB 50%

Thursday 7 June 2018 1.30 to 4.30

COMPUTER SCIENCE Paper 3

*Answer **one** question from each of Sections A, B and C, and **two** questions from Section D.*

*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
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

SECTION A

1 Databases

(*a*) In the database context, what do we mean by *redundant data*? [1 mark]

(*b*) Why might it be a good idea to have redundant data in a database? [2 marks]

(*c*) Why might it be a bad idea to have redundant data in a database? [2 marks]

(*d*) Suppose a database has tables $R(A,\ B)$ and $S(B,\ C)$. Explain how using an *index* could improve performance when joining R and S. Is there a downside to using an index? [4 marks]

(*e*) In SQL, what could be returned when evaluating the following expression?

NOT (a OR (NOT a))

[2 marks]

(*f*) Suppose $R(start, end)$ is a table in a relational database representing arcs in a directed graph. That is, each record $(x, y) \in R$ represents an arc from node x to node y.

(*i*) Write an SQL query that returns the start and end of all 3-hop paths in the directed graph represented by R. Your query should return columns named *start, end*. Each row $(x,\ y)$ in the result of your query should indicate that there exists a path in R

$$x \to z \to u \to y$$

for some nodes z and u. [4 marks]

(*ii*) What is the *transitive closure* of R? Why is this difficult to compute in SQL if we ignore recursive query constructs? [5 marks]

2 Databases

Suppose that we have a relational database with the following tables.

Table	Primary Key
`Movies(mid, title, year)`	`mid`
`People(pid, name)`	`pid`
`Genres(gid, genre)`	`gid`
`ActsIn(pid, mid, role)`	`pid, mid, role`
`HasGenre(gid, mid)`	`gid, mid`

In table `ActsIn`, `pid` is a foreign key into `People` and `mid` is a foreign key into `Movies`. In table `HasGenre`, `mid` is a foreign key into `Movies` and `gid` is a foreign key into `Genres`.

Note that this database is similar to, but not the same as, the examples used in lectures and the database used for practicals.

(a) Suppose that the attribute `role` was not considered part of the key for table `ActsIn`. How would this change your interpretation of the database? [2 marks]

(b) Suppose we replaced the tables `Genres(gid, genre)` and `HasGenre(gid, mid)` with a single table `MovieGenres(mid, genre)`. Would this change what data can be captured in the database? Explain your answer. [2 marks]

(c) Write an SQL query that returns `title`, `mid`, for those movies that are not associated with any genre. (Use the schema at the top of the page, not the possible modifications discussed in (a) or (b).) [4 marks]

(d) Write an SQL query that returns `name`, `pid`, for those people that act in at least one movie associated with the genre `'Drama'`. [5 marks]

(e) Write an SQL query that returns `title`, `mid`, `genre`, for those movies that have `genre` as their only genre. That is, if the query returns the row

`'The Big Hoot'`	`947837`	`'Comedy'`

it means that this movie is associated only with the genre `'Comedy'` and no other genre. [7 marks]

SECTION B

3 Introduction to Graphics

Consider the problem of rendering a scene consisting of spheres graphically using ray tracing.

(*a*) Give a brief overall description of the mathematics underlying the algorithm. Discuss modelling the geometry of individual spheres, formulating the vector equation of a ray, modelling different lighting effects on the surfaces of the spheres, and considering spheres made of refractive and mirrored material. [10 marks]

(*b*) What is meant by *spatial aliasing* and *temporal aliasing* in an image? [3 marks]

(*c*) Describe how super-sampling can be used to reduce spatial aliasing. [4 marks]

(*d*) What is meant by *distributed ray tracing* and when is it used? [3 marks]

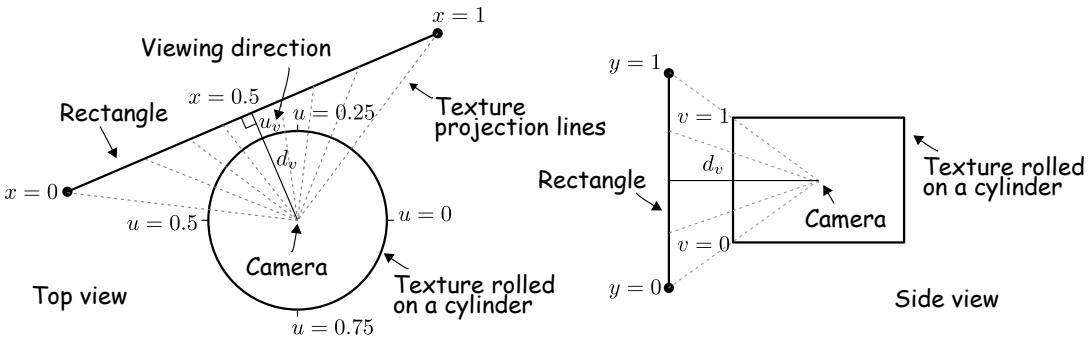
4 Introduction to Graphics

(*a*) When a 2D texture is mapped to a 3D object, the texels can be mapped to areas on the screen that are smaller or larger than the area of a single pixel. Explain what methods are used in OpenGL to ensure the best quality of rendered textures when:

(*i*) Multiple texels are mapped to a single pixel. Explain what image processing operation needs to be performed and what data structure is used in OpenGL to accelerate rendering in this scenario. [5 marks]

(*ii*) There are fewer texels than pixels to be drawn. Explain two interpolation methods available in OpenGL. [5 marks]

(*b*) Your task is to display a portion of a cylindrical panorama (2D image rolled on a cylinder) on a screen. To do this, you need to map a 2D texture from cylindrical coordinates to a rectangle specified by four vertices, as shown in the figure below.



(*i*) Write a formula for computing u, v texture coordinates from object-space coordinates x, y. You are given the distance between the camera and the rectangle, d_v, and the viewing direction u_v in texture coordinates (from 0 to 1). For simplicity, we assume that the full height of the texture is mapped to the full height of the rectangle. Refer to the figure for further details. [6 marks]

(*ii*) Justify whether texture coordinates for mapping a cylinder to a rectangle should be computed in a fragment shader or a vertex shader. [4 marks]

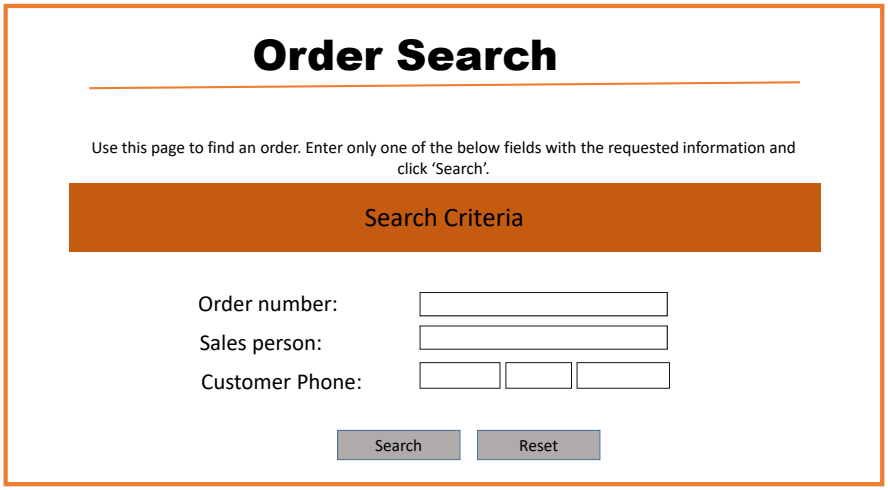
SECTION C

5 Interaction Design

(*a*) During a practical session you were asked to create a working Weather App for a chosen primary stakeholder that works on a desktop or a laptop. This was supposed to be done in three stages: (i) gathering requirements and data, (ii) designing a Low Fidelity (Lo-Fi) prototype, and (iii) implementing a High Fidelity (Hi-Fi) prototype.

Explain the difference between Functional requirements, Non-Functional requirements, and Data requirements. Give examples of Functional, Non-Functional, and Data requirements for the Weather App. [6 marks]

(*b*) Explain what Heuristic Evaluation (HE) is. Evaluate the page provided in the figure below using Heuristic Evaluation by explaining which heuristics have been violated and how, and provide a severity rating with reasons. [6 marks]



(*c*) Explain what GOMS analysis is used to predict. List the steps involved in a GOMS analysis and describe what the GOMS Keystroke-Level Model (KLM) is. Explain three limitations of GOMS analysis. [8 marks]

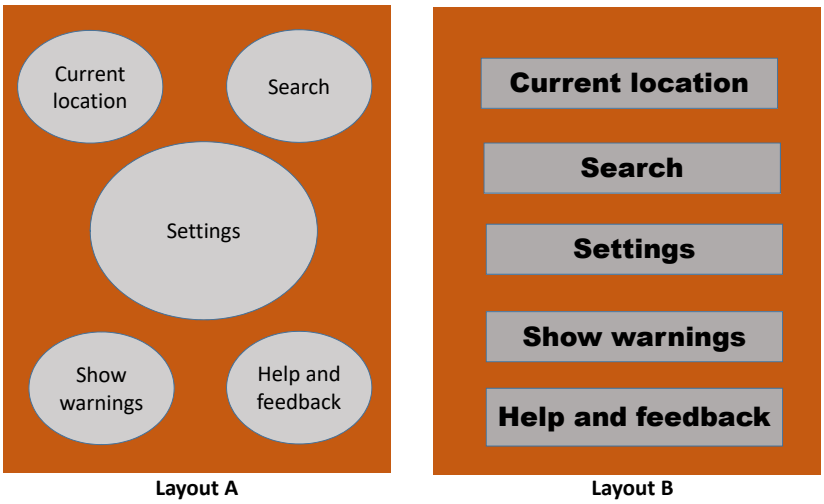
6 Interaction Design

(*a*) During a practical session you were asked to create a working Weather App for a chosen primary stakeholder that works on a desktop or a laptop. This was supposed to be done in three stages: (i) gathering requirements and data, (ii) designing a Low Fidelity (Lo-Fi) prototype, and (iii) implementing a High Fidelity (Hi-Fi) prototype.

Describe the four different kinds of stakeholder in a CUSTOM analysis. For each, provide an example stakeholder from the Weather App. [8 marks]

(*b*) Explain what Lo-Fi Prototyping is and what Wizard-of-Oz Prototyping is. List three different kinds of Lo-Fi prototypes other than Wizard-of-Oz Prototyping. Discuss why Lo-Fi Prototyping is considered to be advantageous over Hi-Fi Prototyping. [6 marks]

(*c*) A company has designed two screen layouts for the home page of a Weather App illustrated in the figure below (Layout A and Layout B). Explain what Hick's Law and Fitts' Law predict. Use Hick's Law and Fitts' Law to compare the two screen layouts (Layout A and Layout B), and explain which layout you think would be faster for users to use and why. [6 marks]



SECTION D

7 Machine Learning and Real-world Data

Some areas of land currently covered by forest had a different previous purpose. An experiment is to be conducted to see whether areas of forest can be automatically classified according to their purpose 50 years ago. There are three categories: meadow, garden and managed woodland. The classification is to be based on the trees currently present: for instance, an area with several apple trees is relatively likely to have been a garden. There are some locations where the true category is known from historical data and the number of trees of each type observed within a fixed distance has been recorded. There is data from 25 meadows, 30 gardens and 45 woodlands. The average number of individual trees at each location is 52.

(*a*) Give formulae for two possible approaches to Naive Bayes classification for this task. [4 marks]

(*b*) How could you derive parameter estimates for use in the Naive Bayes classifiers from this type of data? [4 marks]

(*c*) How would you use the available data to train and test a Naive Bayes classifier? [5 marks]

(*d*) You are now given a large catalogue of tree species with each species manually assigned to zero or more of the categories. Describe a modification to your previous experiment which makes use of this data. [4 marks]

(*e*) You are now given data from 60 new locations. If you use this data solely for evaluation, how would you try and decide whether or not your revised method was an improvement on the original method? [3 marks]

8 Machine Learning and Real-world Data

(*a*) A team plays several seasons with 20 matches in each season. If the team is at home, the probability of winning is 0.7. The probability of winning away is 0.4. No games are draws. Home and away games usually alternate: ignoring the end of the season, there is a probability of 0.75 that a home game is followed by an away game, and also a probability of 0.75 that an away game is followed by a home game. At the start of the season, there is an equal chance of a home or an away game. Show how this information can be partially modelled as a Hidden Markov Model (HMM), treating home and away as hidden states. Give the parameters of the HMM. [6 marks]

(*b*) What aspect of the scenario described is not correctly covered by the HMM? [1 mark]

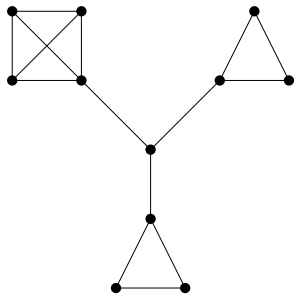
(*c*) If the team's results are 'win, lose, win' at the start of the season, what probability estimate for 'home, away, home' is given by this HMM? [2 marks]

(*d*) You are given a complete record of individual games, including a record of the opponents. The win/loss ratio varies depending on the opponent. Explain how you could use such information to derive the parameters of a more complex HMM (treating home and away as hidden states, as before). [4 marks]

(*e*) It is suggested that you could use an HMM to predict the results of next season's games, since it is known who the opponent will be and whether the match will be at home or away. How might you do this? How successful do you think this would be, compared to predicting whether a sequence of matches were home or away based on a sequence of match results? [7 marks]

9 Machine Learning and Real-world Data

This question concerns undirected graphs which consist of three fully interconnected regions each connected via a single link to a central node. Each fully interconnected region has at least three nodes. See the figure below for an example where the regions are of size 3, 3 and 4.



(*a*) What is the diameter of such a graph? Justify your answer. [1 mark]

(*b*) Which edges are local bridges? [2 marks]

(*c*) What is the betweenness centrality of the central node? Explain your answer. [6 marks]

(*d*) What clusters result if you successively break links in the order given by their edge betweenness centrality (i.e., according to the Newman-Girvan method)? [6 marks]

(*e*) Give an example of a real world situation which might correspond to such a graph. Be explicit about the nature of the nodes and links. Is the clustering result intuitively reasonable for this situation? [5 marks]

END OF PAPER