COMPUTER SCIENCE TRIPOS Part IA

Tuesday 6 June 2017 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

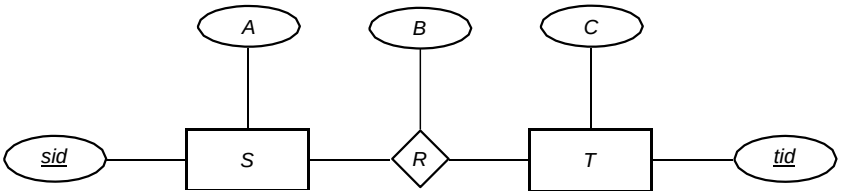
STATIONERY REQUIREMENTS
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

SECTION A

1 Databases

(a) Consider the following Entity-Relationship (ER) diagram.



Suppose we wish to implement this diagram in a relational database using three tables, $S(\underline{sid}, A)$, $T(\underline{tid}, C)$, and $R(\cdots)$. Describe the schema you would use for R depending on the cardinality of the relationship.

(i) When R is a many-to-many relationship between S and T. [2 marks]

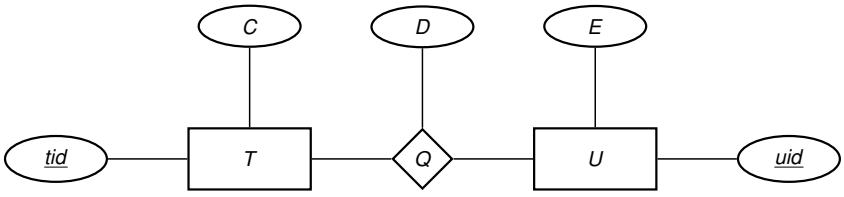
(ii) When R is a one-to-many relationship between S and T. [2 marks]

(iii) When R is a many-to-one relationship between S and T. [2 marks]

(iv) When R is a one-to-one relationship between S and T. [2 marks]

(b) Suppose R is a many-to-one relationship. Rather than implementing a new table for R, can we modify one of the tables representing S or T to implement this relationship? Discuss the advantages and disadvantages of such a representation. [4 marks]

(c) Suppose that we add the following diagram to our ER model.



Note that this implicitly defines a relationship between S and U resulting from the composition of relationships R and Q. Discuss the difficulties that you might encounter in attempting to implement this derived relationship directly in a table W. For example, would the results of evaluating this SQL

```
select sid, tid, B, D
from R
join Q on R.tid = Q.tid
```

always be equivalent to the contents of such a W? [8 marks]

2 Databases

Consider the following three tables.

$S(\underline{sid},\ A)$

sid	A
s1	a1
s2	a2
s3	a3

$R(\underline{sid},\ \underline{tid},\ B)$

sid	tid	B
s1	t1	b1
s1	t2	b2
s2	t1	b4
s2	t3	b5

$T(\underline{tid},\ C)$

tid	C
t1	c1
t2	c2
t3	c3
t4	c4

(a) Represent the information in these three tables in a single table, using NULL values where needed. [4 marks]

(b) Represent the information in these three tables as three JSON objects, each associated with one of the values of the *sid* key. Is any information lost? [4 marks]

(c) Represent the information in these three tables as four JSON objects, each associated with one of the values of the *tid* key. Is any information lost? [4 marks]

(d) We now have three distinct ways of representing the same information (the original tables, one big table, and the collection of JSON objects from parts (b) and (c)). Carefully compare and contrast these approaches and discuss their related advantages and disadvantages. [8 marks]

SECTION B

3 Graphics

Consider display technologies for hand-held devices.

(*a*) Explain the principles of operation of each of the following.
Note: You may illustrate your answers with a diagram.

(*i*) Liquid crystal displays [5 marks]

(*ii*) Electrophoretic (electronic paper) displays [5 marks]

(*b*) Compare and contrast their characteristics. [6 marks]

(*c*) Explain how liquid crystal and electrophoretic displays can show coloured images. [4 marks]

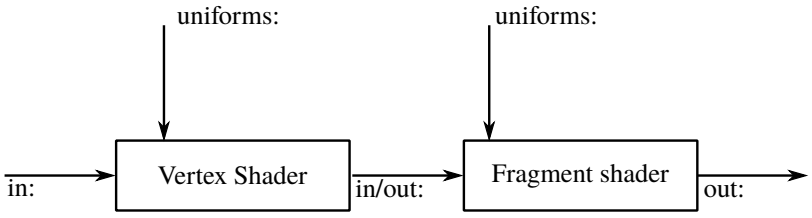
4 Graphics

Consider rendering a triangular mesh using OpenGL. A uniform material is used for the entire mesh and the reflection model of the material consists of ambient, diffuse and specular components. There are two point light sources in the scene. Given these assumptions, answer the following questions:

(*a*) Gourand and Phong shading are two different methods of interpolating colours between vertices. Explain how each method interpolates colours. [5 marks]

(*b*) Discuss the trade-offs in terms of quality and computational costs for Phong and Gourand shading. Assume that the number of rendered pixels is much larger than the number of vertices. What kind of artefacts can one of the methods produce and what is the reason for those artefacts? [7 marks]

(*c*) For each of Gourand and Phong shading, explain how you would implement each shading method using vertex and fragment shaders in OpenGL. Complete the diagram shown below by listing all inputs, outputs and uniforms for each shader. Then, explain what is computed in each shader in the case of both shading methods. There is no need to write equations or code, but you may include them if it helps your explanation.



[8 marks]

SECTION C

5 Interaction Design

Your company has been asked to design and implement an interface called GestureDriver which is envisioned to be a remote driving interface based on visual gesturing and speech. Hand motions will be tracked based on colour vision, and the system will classify the gestures using a simple geometric model. The gestures will be mapped into six motion commands: start, right, left, forward, reverse, and stop. These commands can also be provided via speech and transmitted to the remote vehicle. The major requirements are that the system displays the current location of the vehicle to the user, monitors location change, captures the motion and vocal command of the user in real time, and moves the vehicle based on the received command.

(*a*) One of the Principles of Good Design is to 'balance automated and human control'. Give four examples and explain how this can be achieved for the context provided above. [4 marks]

(*b*) Describe, with an example, each of the five different interaction spaces beyond the desktop as classified by Winograd. Briefly discuss which one(s) would be more appropriate to use for creating this application, and why. [6 marks]

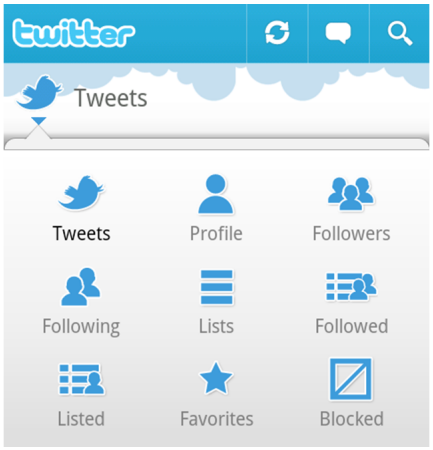
(*c*) Assuming that the initial system status is OFF, draw a task model for the user using the GestureDriver interface with the goal of reversing the vehicle they are operating remotely. The task model should capture activities at multiple levels starting with at least three activities. Explain any sequential dependencies you identify. [8 marks]

(*d*) Describe one generic problem that task analysis can help identify, and give an example of how this can be done. [2 marks]

6 Interaction Design

(*a*) Describe three problems that affect the data and requirements gathering process, and provide one solution for each problem. [3 marks]

(*b*) State and briefly describe each of the six principles of Gestalt psychology that have implications for user interface design. Explain how each of the six principles of Gestalt psychology have been employed in the design of the interface given in the figure below. [12 marks]



(*c*) A company has just completed the heuristic evaluation of a digital music player, and identified the problem areas provided in the list below. Indicate which usability principle or principles each problem violates for five of the following problems. [*Note:* it is possible that each problem violates more than one usability principle, and that the same principle is violated multiple times for different problems.]

P1. Inconsistencies between menus and buttons.

P2. Some language does not correspond with user terminology.

P3. There are buttons that the user may not realize are buttons.

P4. Not all buttons have tooltips.

P5. There are inconsistencies with Windows operating system standards.

P6. Undo commands basically unsupported.

P7. Help content uses different terminology from application.

P8. System does not always provide user with enough information about the task being performed. [5 marks]

SECTION D

7 Machine Learning and Real-world Data

Suppose that a very large collection of documents describing the University of Cambridge has been collected. Your task is to build a classifier which assigns sentiment to each document, using the classes: *positive*, *negative* and *neutral*.

(*a*) There is no ground truth sentiment associated with the documents, but 200 have been manually classified by three human annotators. Fleiss' Kappa is 0.65 for this set. Explain what Kappa is and outline how it is calculated. You do not need to state the full formula for Kappa. [4 marks]

(*b*) Given the limited amount of annotated data, you decide to classify the documents using a standard sentiment lexicon. Explain how you would perform an experiment to do this. [5 marks]

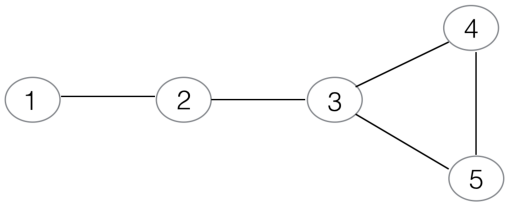
(*c*) (*i*) How would you evaluate the results of the system you have described in your answer to part (*b*) using the annotated data? Give details of the evaluation metrics you would use. [4 marks]

(*ii*) If the primary objective were to identify the documents with negative sentiment, how might your proposed evaluation change? [2 marks]

(*d*) It is suggested to you that the classes automatically assigned by the sentiment lexicon approach could be used to provide training data for a Naive Bayes classifier. Could such a classifier perform better than the sentiment lexicon classifier which provided the decisions it was trained on? Explain your answer. [5 marks]

8 Machine Learning and Real-world Data

Consider the following graph of users in a social network.



(*a*) Give definitions of the following quantities, and state their values for nodes 1, 2 and 3:

(*i*) Degree [1 mark]

(*ii*) Clustering Coefficient [2 marks]

(*iii*) Betweenness Centrality [2 marks]

(*b*) (*i*) Which property of nodes in a network do clustering coefficient and betweenness centrality attempt to model? Give examples from naturally occurring networks. [4 marks]

(*ii*) In the network depicted above, do the betweenness centrality values you calculated correspond to your intuition about the function of nodes 1, 2, and 3? If not, which factors of the formula are responsible for the divergence? [3 marks]

(*c*) Suppose the link between 3 and 4 is removed and a link between 2 and 4 is added. How does this affect the betweenness centrality of 2 and 3, and why? [3 marks]

(*d*) Some pairs of nodes in the new network from part (*c*) now have more than one shortest path that connects them. Identify these paths, and explain how their existence affects the calculation of betweenness centrality. [2 marks]

(*e*) Professor Miller claims that it is possible for a particular pair of nodes in a graph to be connected by a number of shortest paths which is exponential with respect to the number of nodes in the graph. Give an example where this is the case, or prove that it cannot be the case. [3 marks]

9 Machine Learning and Real-world Data

Hidden Markov Models (HMM) can be used to find names in text. In the following HMM created for this purpose, the two emitting states are q_1 = `i` (for "inside a name") and q_2 = `o` (for "outside a name"). Each word in the training data is labelled with either `i` or `o`. There are two sequences in the training data, as follows:

```
today may bakes a nice cake
 o      i   o   o  o    o

peter bakes and mary bakes may like sue
  i    i     o    i    i    o   o    i
```

(*a*) Give the general formula for estimating transition probabilities from training data. Provide the full transition matrix A for this HMM based on the training data shown. [6 marks]

(*b*) Give the general formula for calculating emission probabilities from training data, and calculate the emission probabilities $P(may|o)$, $P(may|i)$, $P(bakes|o)$, $P(bakes|i)$. [3 marks]

(*c*) An HMM trained with the above training observations is exposed to the following test observation:

```
may bakes
```

Which probabilities does the HMM assign to the following two interpretations?

(*i*) `may` is a name, and `bakes` is not [2 marks]

(*ii*) `bakes` is a name, and `may` is not [2 marks]

(*d*) The first training observation is now replaced with:

```
today peter bakes a nice cake
```

How does this change your answers to part (*c*)? [2 marks]

(*e*) A comparable situation to part (*d*) can arise even with substantial amounts of training data. Describe why this is a problem and indicate a solution to it. [5 marks]

END OF PAPER