CST0
COMPUTER SCIENCE TRIPOS Part IA

Thursday 9 June 2022 14:00 to 17:00 BST

COMPUTER SCIENCE Paper 3

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

*Submit each question answer in a **separate** PDF. As the file name, use your candidate number, paper and question number (e.g., `1234A-p3-q6.pdf`). Also write your candidate number, paper and question number at the start of each PDF.*

You must follow the official form and conduct instructions for this online examination

SECTION A

1 Databases

You wish to query a database which is a subset of the IMDb Internet Movie Database. [*Note:* The first course practical used such a database.] Recall that the database schema has table `movies` with key `movie_id`, the table `people` with key `person_id`, and the table `genres` with key `genre_id`. The table `has_genre` implements a relationship between `movies` and `genres` and has the key `(movie_id, genre_id)`. The table `plays_role` implements a relationship between `movies` and `people` and has the key `(movie_id, person_id, role)`.

(*a*) Write an SQL query to return the number of movies that are romantic comedies.

[6 marks]

(*b*) Complete the following SQL so that it returns records of the form

```
pid1, pid2, movie_id
```

where `pid1` and `pid2` are identifiers of co-actors with roles in the romantic comedy with identifier `movie_id`. This should be a symmetric table so that if `pid1, pid2, m` is in the result, then so should be `pid2, pid1, m`. However, it should not include records where `pid1` and `pid2` are equal.

```
select R1.person_id as pid1,
       R2.person_id as pid2,
       M.movie_id as movie_id
from .... your code goes here ....
```

[7 marks]

(*c*) Complete the following SQL so that it returns records of the form

```
name1, title1, name2, title2, name3
```

that can be interpreted as follows:

- Actors `name1` and `name2` are co-actors in a romantic comedy `title1`.
- Actors `name2` and `name3` are co-actors in a romantic comedy `title2`.
- However, neither actor `name1` has a role in the movie associated with `title2` and `name3`, nor does actor `name3` have a role in the movie associated with `title1` and `name1`.

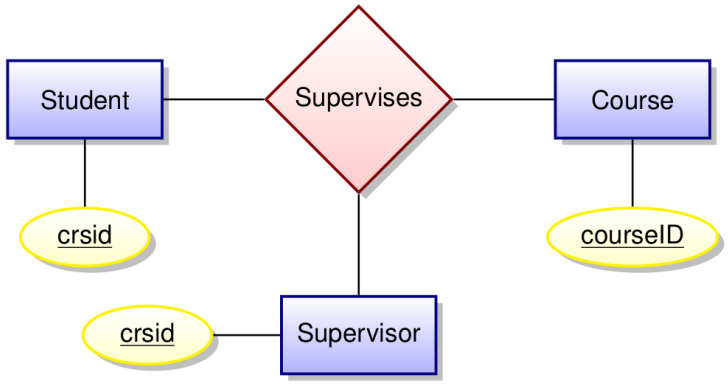
```
select P1.name as name1, M1.title as title1,
       P2.name as name2, M2.title as title2,
       P3.name as name3
from .... your code goes here ....
```

[7 marks]

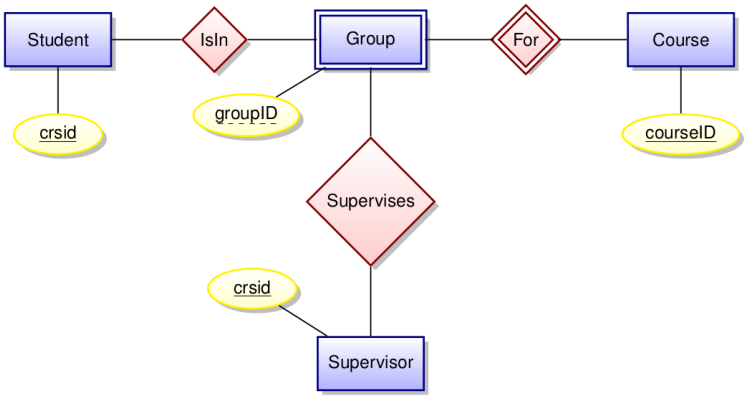
2 Databases

(a) Part II supervisions are organised within the department rather than colleges. The department intends to implement a relational database application to track the allocation of supervisors to students. As a preliminary step it has asked two data modellers each to provide an Entity-Relationship for this task. The modellers delivered the two different ER diagrams below. Note that both models are incomplete in that many obvious attributes are missing as well as the cardinality constraints on relationships. However, argue that diagram M_2 represents a better initial model than diagram M_1. [5 marks]

ER model M_1:



ER model M_2:

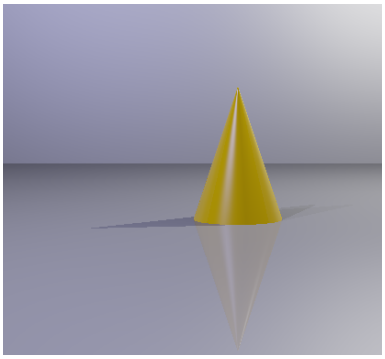


(b) Present one way of implementing model M_2 in a relational database. Note that you first have to determine reasonable cardinalities for the relationships in M_2. Justify your choices. [8 marks]

(c) Using your relational implementations from Part (b), write an SQL query that returns records of the form `crsid, courseid, groupid` where `crsid` is the id of a student in a supervision group for course with id `courseid`, and the supervision group id is `groupid`. The `groupid` column should contain NULL if the student is not in any supervision group for the associated course. [7 marks]

SECTION B

3 Introduction to Graphics



A cone, shown in the figure above, is given by its implicit equation:

$$x^2 + z^2 - \left(r - \frac{r}{h}y\right)^2 = 0 \tag{1}$$

where r is the radius of the base and h is its height. The centre of the base is at the origin and the apex lies on the y-axis.

(a) Derive an equation for the intersection of ray $(O, D) = ([O_x\ O_y\ O_z]^T, [D_x\ D_y\ D_z]^T)$ with:

(i) the base of the cone; [3 marks]

(ii) the sides of the cone. You may leave the equation for the sides in the quadratic form. [6 marks]

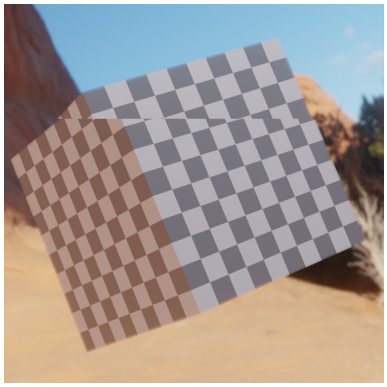
You must use the implicit representation from the equation above.

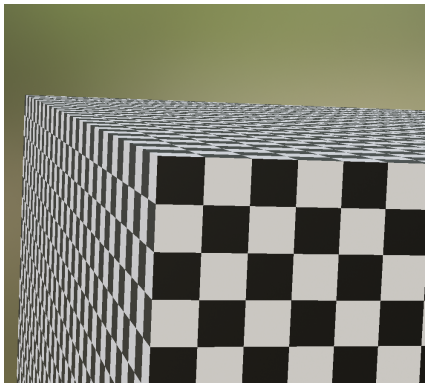
(b) Write the equation for the normal at the surface of the side of the cone at the coordinates (x, y, z). [4 marks]

(c) You want to rotate the cone about x-axis by angle α, then about the z-axis by angle β and finally translate it by vector $t = [t_x\ t_y\ t_z]^T$. Find the point of intersection of the ray from Part (a) with the transformed cone. You may reuse the equations from Part (a) of the question and leave the transformation matrices as $R_x(\alpha)$, $R_z(\beta)$ and $T(t)$ without writing down their contents. [7 marks]

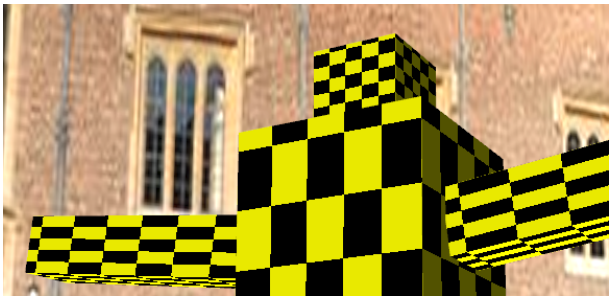
4 Introduction to Graphics

(a) The two screenshots below show two different OpenGL rendering artifacts. Identify the artifacts, explain their cause and how to avoid them. [6 marks]





(b) The simple GLSL shader below puts a yellow diffuse material on an object:



```
in vec3 N;
in vec2 tex_uv;
in vec3 wc_frag_pos;
out vec3 color;


void main() {
    const vec3 I_a = vec3(1, 1, 1) * .008;
    const float k_d = 0.4;
    const vec3 I = vec3(1, 1, 0.9);
    const vec3 L = vec3(1, -3, 1);
    vec3 C_d = vec3(1, 1, 0);
    vec3 linear_color = C_d*I_a + C_d*k_d*I*max(0, dot(N, L));
    color = tonemap(linear_color);
}
```

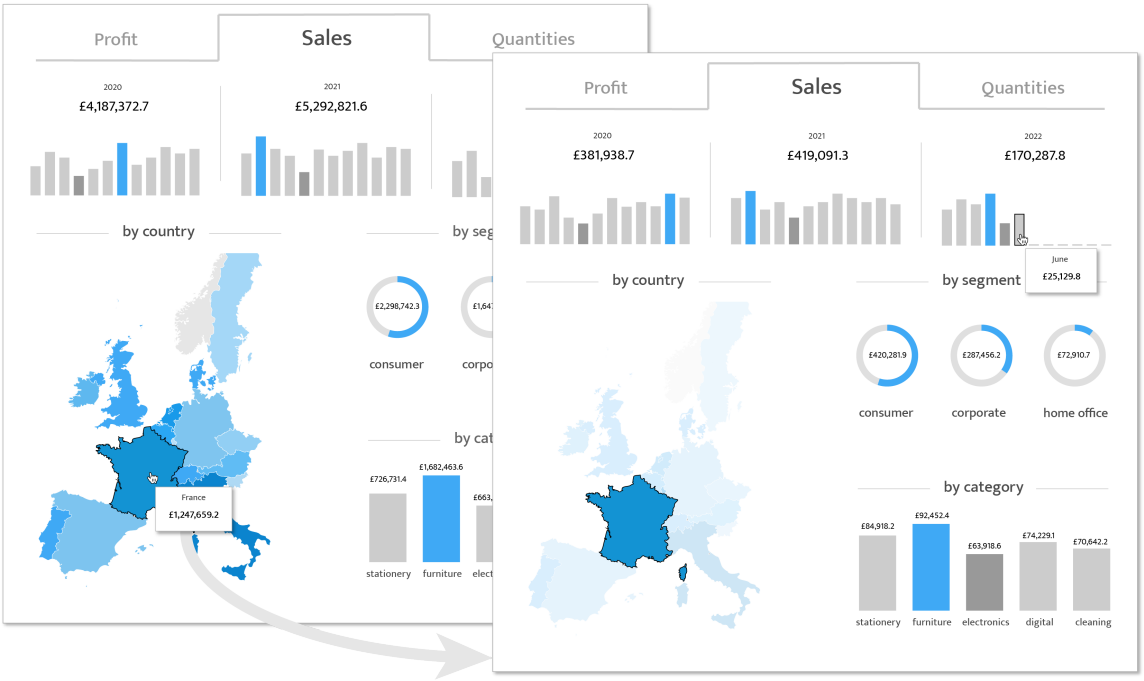
Add a few lines of code that generate a black-and-yellow checkerboard pattern as shown in the image next to the code. Assume that the correct uv coordinates have been passed from the vertex shader but no texture has been specified. You may use an arbitrary scale of the pattern on the surface. [7 marks]

(c) You measure red, green and blue primaries of a standard dynamic-range display to be $[X_r\ Y_r\ Z_r]$, $[X_g\ Y_g\ Z_g]$, $[X_b\ Y_b\ Z_b]$. The gamma of the display is γ. Write the equation for mapping from linear colour $[R_{\text{in}}\ G_{\text{in}}\ B_{\text{in}}]$ in the ITU-R 2020 colour space to the display-encoded RGB in the native colour space of the display. You are provided with the matrix $M_{\text{R2020toXYZ}}$ mapping from ITU-R 2020 to XYZ colour space. [7 marks]

SECTION C

5 Interaction Design

You are part of the data analytics and visualisation team at a large superstore chain. Below is the latest dashboard that your team has built to show yearly profits, sales, and quantities sold.



(*a*) Identify and describe three gestalt principles and explain how these have been used in the design of the dashboard. [6 marks]

(*b*) Before making the dashboard available to the business team, your team wants to conduct a Cognitive Walkthrough to identify usability issues. The typical user is a business manager who is overseeing a number of countries. The task to be tested is to "Find the latest monthly sales total for France".

Write an instruction sheet for first-time evaluators. Your instructions should include a description of what Cognitive Walkthrough is, the action sequence for achieving the task to be tested, and what the evaluators should record for each action. [5 marks]

(*c*) The business team have requested a version suitable for mobile devices. Identify three aspects that need redesign due to the differences between desktop and mobile devices. For each aspect, describe how interaction on desktop and mobile devices differs and sketch one potential design solution. [9 marks]

6 Interaction Design

As part of the government's response to the COVID-19 pandemic, your design agency has been contracted by the Ministry of Health to design an application for reporting test results and notifying members of the population when they need to isolate.

(*a*) What user research methods would you use for data gathering before designing? List two methods that you would recommend using, why you think they would be suitable for this project, and what data you expect to collect with each of the methods. [6 marks]

(*b*) Following data gathering using the methods listed above, your team has conducted stakeholder and requirements analyses. List and describe the identified stakeholders. Identify three key requirements that the application must meet. [6 marks]

(*c*) Propose and motivate one design principle emerging from the stakeholder and requirements analyses above which the design of the application should follow. [4 marks]

(*d*) Discuss one potential trade-off between usability and data security & privacy. [4 marks]

SECTION D

7 Machine Learning and Real-world Data

A new type of fraud that has been going on for the past 48 hours has been detected by a bank. The bank has manually established for each of the 120,000 transactions that occurred in this time whether it was affected by the fraud; this was the case in 14 transactions. The bank plans to deploy an automated fraud-detection system that treats the problem as binary classification.

(a) Two competing systems are to hand for the task: System A declares 6 transactions as fraud, of which 5 were indeed fraudulent. System B declares 32 transactions as fraud, out of which 12 were fraudulent. There are 5 transactions for which both systems declare fraud, out of which 4 are fraudulent.

(i) For each system, give the number of false negatives, true negatives, false positives and true positives. [2 marks]

(ii) Define and calculate the accuracy on the given data for both systems. [2 marks]

(iii) Define and calculate the precision and recall of fraud for both systems. [3 marks]

(b) The bank director decides to deploy the system with the higher accuracy, if it turns out to be significantly better than the other system. She asks you to perform a sign test to determine if this is so.

(i) Describe how you would proceed in principle. What is your null hypothesis? What are parameters N and p in the relevant formula? [*Note:* It is not necessary to give the formula.] [2 marks]

(ii) In how many transactions does System A beat System B and vice versa? In how many transactions do they perform identically? [4 marks]

(iii) Does accuracy, in combination with the sign test from Part (b)(i), adequately distinguish between the systems? [2 marks]

(iv) Is there a more meaningful comparison of the two systems you can offer? Can you still perform a significance test on the metric of your choice? If so, how? If not, why not? [5 marks]

8 Machine Learning and Real-world Data

A local government wants to use the COVID-19 test positivity rate (i.e. the proportion of tests with a positive result) to estimate the number of COVID-19 infections in the community. In the data collected, the positivity rate is labelled with one of three levels of positivity (`+`, `++` and `+++`) and the number of people infected are categorised as high (H), medium (M) or low (L).

timestep	1	2	3	4	5	6	7
infections	L	M	H	H	H	M	M
positivity	+	++	++	++	+++	++	++

They decided to use a first-order hidden Markov model (HMM), modelling the infections as the hidden states and the positivity as the observations.

(*a*) Define and estimate the components of an appropriate HMM for this application, without smoothing. [4 marks]

(*b*) What assumptions are implicit with the use of an HMM? Are they appropriate in the context of this application? [4 marks]

(*c*) You are in timestep 7 and you now observe the following positivity rates, but you do not know the number of infections:

timestep	8	9	10
positivity	+++	++	++

Predict the number of infections for each timestep using the Viterbi algorithm, showing the equations and calculations you make. [8 marks]

(*d*) Briefly describe two shortcomings of the HMM developed for predicting the number of infections. [4 marks]

[*Note:* This version fixes a typesetting mistake that had appeared in the exam.]

9 Machine Learning and Real-world Data

An online brand-safety company has as its goal to warn its clients when there is a problem with their brand name on social media. To this end, it wants to develop a classifier to determine whether a brand is being attacked on social-media platforms. To develop this classifier, the company collected 5000 social-media posts: 100 posts referring to each of 50 brands. For each brand, company employees read all 100 posts, and hand-annotated the brand with one of the two labels, *attacked* and *safe*; this resulted in 10 brands being labelled *attacked* and 40 brands labelled *safe*. Your task is to develop a naive Bayes classifier that uses the text of the posts as features.

(a) Give the equations for a naive Bayes classifier for the task of determining whether a brand is safe or being attacked. [2 marks]

(b) How would you split the data into training and testing? Justify your choice. [2 marks]

(c) Here are some posts for two brands used for training:

```
BRAND  LABEL  POSTS
GStuff safe     Awesome products as always from GStuff!
GStuff safe     Fantastic customer support from GStuff!
GStuff safe     Awesome GStuff product, I wish it were cheaper
AThing attacked Not the best experience from AThing
AThing attacked Will not buy AThing again! Terrible performance
AThing attacked Terrible performance from AThing, avoid
```

(i) What features do you expect your naive Bayes classifier to consider important? Give three examples for the 10 *attacked* brands and also three examples for the 40 *safe* brands. How well do you expect each of these features to generalize? [6 marks]

(ii) Based on your observations in Part (c)(i), suggest and justify two changes to feature extraction to improve generalization. [4 marks]

(d) You claim that a future product will enable you to warn clients about attacks before any competitor does. Give three modifications to your classifier and the evaluation setup that might help you achieve this. [3 marks]

(e) Consider changing your approach to a classifier operating at the post level, i.e. classifying media posts instead of brands. There are advantages and disadvantages of doing so. Give three. [3 marks]

END OF PAPER