CST0+CST1
COMPUTER SCIENCE TRIPOS Part IA 75%, Part IB 50%

Monday 7 June 2021 11:30 to 14:30 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

The following tables are part of a library's database system.

Book(**book_id**, title, number_owned, number_borrowed)
Person(**person_id**, name, address)
Borrowed(**person_id, book_id**, number)

The primary keys of each table are in **bold**. In the table Book the column `number_owned` is the number of copies of the book owned by the library, while the column `number_borrowed` is the number of copies currently out on loan. In table Borrowed the column `person_id` is a foreign key into the Person table, the column `book_id` is a foreign key into the Book table. The column `number` is the number of copies of the book borrowed by the associated person. (This library is used by primary school teachers who frequently check out many copies of a book for the use in their classes.)

If the database is internally consistent, then the column `number_borrowed` is redundant information that can be computed from the actual number borrowed, and this can be derived from the Borrowed table.

(*a*) Write an SQL query that checks the internal consistency of this database. It should return records of the form

```
(book_id, number_borrowed, actual_number_borrowed)
```

only for those books where `number_borrowed` and `actual_number_borrowed` are not equal. That is, if the database is consistent the query will return no records. [5 marks]

(*b*) Your job is to redesign this schema so that there is no need for such consistency checks. The first step is to design an Entity-Relationship model. You will do this by introducing a new entity called Copy_Of. Each copy of a book owned by the library will be associated with a unique member of the Copy_Of entity.

Design an Entity-Relationship diagram based on this idea and argue that cardinality constraints will ensure that the database is internally consistent. [5 marks]

(*c*) Discuss at least two options for implementing your ER model in an SQL database. [5 marks]

(*d*) Using one of your relational implementations from the previous part, write an SQL query that reproduces the contents of the Book table from the original design. That is, write an SQL query that returns records of the form

```
(book_id, title, number_owned, number_borrowed).
```

[5 marks]

2 Databases

This question involves the relational movie database used in our SQL practical.

In each part you are given the result of an SQL query together with a possibly incorrect conclusion drawn from this result.

In each case your task is to argue for or against the conclusion. You must clearly justify your reasoning. If the SQL query can be corrected, then do so.

(*a*) The following query returns 1422.

Conclusion: Our database contains information on 1422 directors.

```
select count(*)
from has_position
where position = 'director';
```

[6 marks]

(*b*) The following query returns these records:

```
PERSON_ID  NAME          POSITION  TOTAL
---------  ------------  --------  -----
nm0498278  Stan Lee      writer       15
```

Conclusion: Stan Lee did not produce any of the movies in our database.

```
select person_id, name, position, count(*) as total
from has_position as hp
join people as p on p.person_id = hp.person_id
where position <> 'actor'
  and name = 'Stan Lee'
group by person_id, name, position
```

[6 marks]

[**continued . . .**]

(*c*) The following query attempts to return records `(role, year, total)` where Jennifer Lawrence plays the same `role` during the `year` a `total` number of times in different movies. The query returns these records:

```
ROLE              YEAR  TOTAL
----------------  ----  -----
Tiffany            2012      1
Mystique           2011      1
Raven              2011      1
Aurora Lane        2016      1
Katniss Everdeen   2012      2
Ree                2010      1
Rosalyn Rosenfeld  2013      1
Katniss Everdeen   2013      2
```

Conclusion: Jennifer Lawrence played Katniss Everdeen in two movies in 2012.

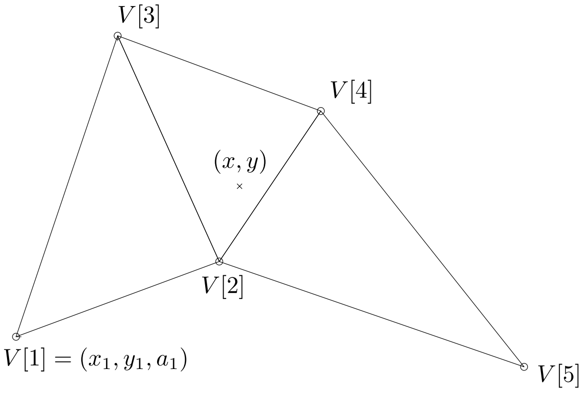
```
select r1.role, m1.year, count(*) as total
from plays_role as r1
join plays_role as r2 on r2.person_id = r1.person_id
join movies as m1 on m1.movie_id = r1.movie_id
join movies as m2 on m2.movie_id = r2.movie_id
join people as p on p.person_id = r1.person_id
where p.name = 'Jennifer Lawrence'
      and r1.role = r2.role
group by r1.role, m1.year;
```

[8 marks]

SECTION B

3 Introduction to Graphics

You are provided with a 2D triangle mesh defined by a set of vertices $V[k] = (x_k, y_k, a_k)$ for $k = 1, \ldots, N$, and a triangle index table T of dimension $M \times 3$, where M is the number of triangles. x_k and y_k are the coordinates of vertex k and a_k is its scalar attribute. An example of such a triangle mesh is shown below.



(*a*) Write down the triangle index table of the mesh shown in the figure above. Ensure that all triangles are forward-facing. [4 marks]

(*b*) Write pseudocode for a function $a = \text{lookup\_a}(x, y, V, T)$, which returns the value of the linearly interpolated attribute at the point (x, y) when the point lies on the mesh and -1 otherwise. Use square brackets to index vertex ($V[i]$) and triangle ($T[i, j]$) tables. The pseudocode should include the formulas needed to compute the interpolated attribute value and to check whether the point is inside the triangles. [10 marks]

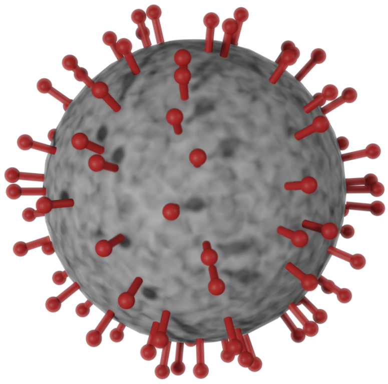
(*c*) Suppose that now vertices also include a depth, so that $V[k] = (x_k, y_k, z_k, a_k)$, and triangles overlap and occlude one another. How do you need to modify the pseudocode to return the attribute of the visible triangle that has the lowest z-value at a given point? Due to memory limitations, you cannot use the Z-buffer algorithm. [6 marks]

4 Introduction to Graphics

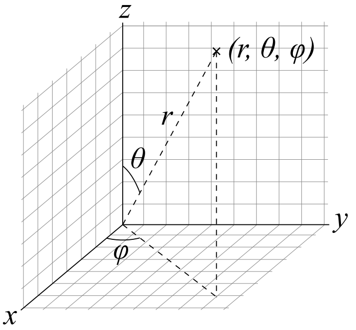
Your task is to create a simple visualization of SARS-CoV-2 using only two primitives:

- a sphere of radius 1, and
- a cylinder of a height 2, its base of radius 1, and the main axis aligned with OZ,

both centred at the origin. An example of such a visualization is shown in Fig. 1 below.



(Fig. 1)



(Fig. 2)

(a) Draw a scene graph for the SARS-CoV-2 model, shown in Fig. 1, in which the main shape is a sphere and each spike consists of a cylinder and a sphere. Use the hierarchy of the primitives (sphere — cylinder — sphere) so that the entire object can be animated by transforming the main shape. [4 marks]

(b) Provide transformation matrices for each node of the graph. You do not need to provide the results of the matrix multiplication. Assume that you have a list of N spherical coordinates (ϕ_k, θ_k), for $k = 1, \ldots, N$, which indicate the positions of the spikes. Use the coordinates as shown in Fig. 2. The main body has a radius of 1, the cylinder of the spike has a length of 0.1, a radius of 0.025 and the sphere of the spike has a radius of 0.05 with the centre at the base of the cylinder. [10 marks]

(c) How can you randomly generate the spherical coordinates (ϕ_k,θ_k) of the spikes so that they are (a) evenly distributed over the sphere and (b) not clustered together (two or more spikes are not too close to each other)? Write pseudocode for generating (ϕ_k,θ_k). [6 marks]

SECTION C

5 Interaction Design

Nowadays, timepieces (such as clocks, wristwatches, etc.) have a variety of functions. They not only tell the time and date but they can speak to you, remind you when it's time to do something, and provide a light in the dark, among other things. Mostly, the interface for these devices, however, shows the time in one of two basic ways: as a digital number such as 23:40 or through an analog display with two or three hands – one to represent the hour, one for the minutes, and one for the seconds.

In this question, we ask you to design a new timepiece for your own use. This could be in the form of a wristwatch, a mantelpiece clock, an electronic clock, or any other kind of timepiece you fancy.

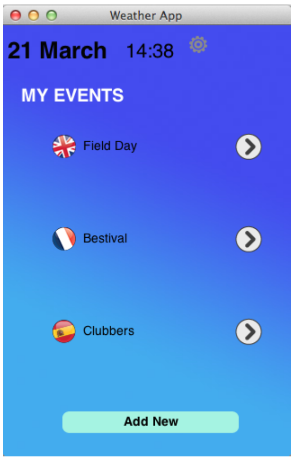
(*a*) Think about the interactive product you are designing: describe what you want it to do for you. Write a list of functional and non-functional requirements. [4 marks]

(*b*) Sketch out an initial low-fidelity prototype for the timepiece and develop at least two distinct alternatives that both meet your set of requirements listed above. [8 marks]

(*c*) Nielsen's heuristics used for Heuristic Evaluation are: (1) visibility of system status, (2) match between system and real world, (3) user control and freedom, (4) consistency and standards, (5) error prevention, (6) recognition rather than recall, (7) flexibility and efficiency of use, (8) aesthetic and minimalist design, (9) help users recognize and recover from errors, and (10) help and documentation.

Evaluate the two low-fidelity prototypes using Heuristic Evaluation. [8 marks]

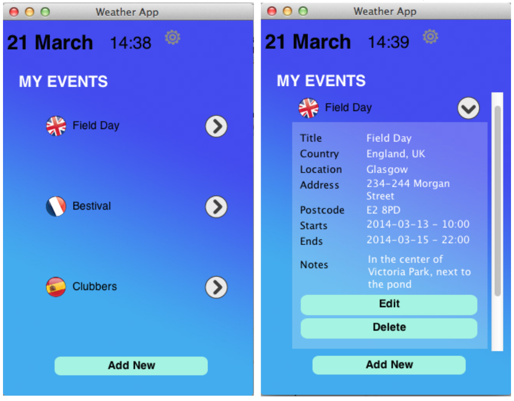
6 Interaction Design



A weather app, whose main screen is shown in the figure above, has been created specifically for event organizers as primary stakeholders.

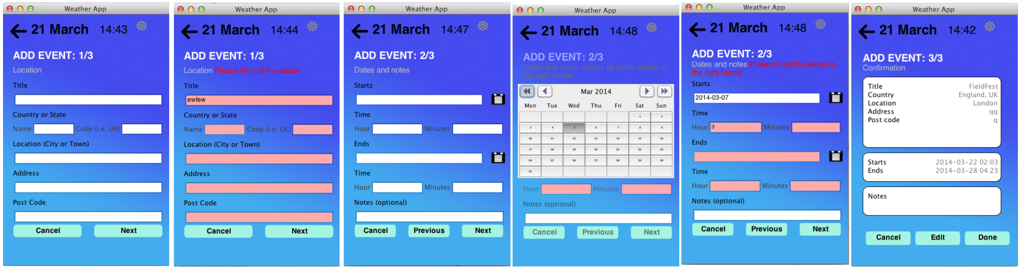
(*a*) Evaluate the provided low-fidelity prototype using Cognitive Walkthrough and the visualised screens (if you need to, you may come up with your own assumptions regarding the primary stakeholders), for the following tasks:

(*i*) (starting from the main screen) View details of the event called Field Day



Title Field Day
Country England, UK
Location Glasgow
Address 234–244 Morgan Street
Postcode E2 8PD
Starts 2014–03–13 – 10:00
Ends 2014–03–15 – 22:00
Notes In the center of Victoria Park, next to the pond

(*ii*) (starting from the main screen) Add a new event



Title: ewfew

Mon	Tue	Wed	Thu	Fri	Sat	Sun
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

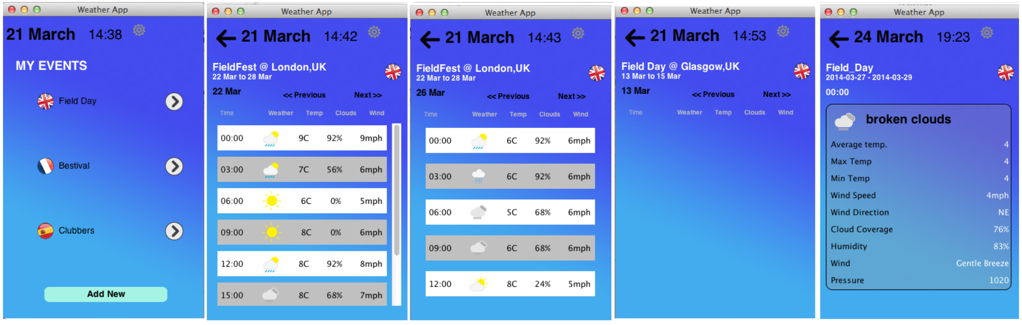
Starts: 2014-03-07
Hour: 3

Title FieldFest
Country England, UK
Location London
Address qq
Post code q

Starts 2014–03–22 02:03
Ends 2014–03–28 04:23

[continued . . .]

(*iii*) (starting from the main screen) View detailed weather forecast for Field Day



Weather App

21 March 14:38

MY EVENTS

Field Day

Bestival

Clubbers

Add New

Weather App

21 March 14:42

FieldFest @ London,UK
22 Mar to 28 Mar

22 Mar << Previous Next >>

Time	Weather	Temp	Clouds	Wind
00:00		9C	92%	9mph
03:00		7C	56%	6mph
06:00		6C	0%	5mph
09:00		8C	0%	6mph
12:00		8C	92%	8mph
15:00		8C	68%	7mph

Weather App

21 March 14:43

FieldFest @ London,UK
22 Mar to 28 Mar

26 Mar << Previous Next >>

Time	Weather	Temp	Clouds	Wind
00:00		6C	92%	6mph
03:00		6C	92%	6mph
06:00		5C	68%	6mph
09:00		6C	68%	6mph
12:00		8C	24%	5mph

Weather App

21 March 14:53

Field Day @ Glasgow,UK
13 Mar to 15 Mar

13 Mar << Previous Next >>

Time	Weather	Temp	Clouds	Wind

Weather App

24 March 19:23

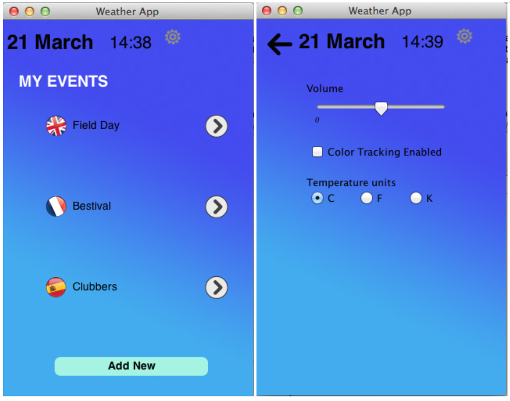
Field_Day
2014-03-27 - 2014-03-29

00:00

broken clouds

Average temp.	4
Max Temp	4
Min Temp	4
Wind Speed	4mph
Wind Direction	NE
Cloud Coverage	76%
Humidity	83%
Wind	Gentle Breeze
Pressure	1020

(*iv*) (starting from the main screen) Change the units of measurement for the temperature information provided



Weather App

21 March 14:38

MY EVENTS

Field Day

Bestival

Clubbers

Add New

Weather App

21 March 14:39

Volume

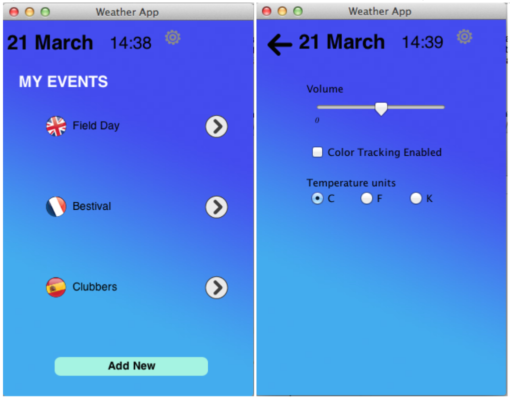
0

□ Color Tracking Enabled

Temperature units

● C ○ F ○ K

(*v*) (starting from the main screen) Modify the volume setting to 150



Weather App

21 March 14:38

MY EVENTS

Field Day

Bestival

Clubbers

Add New

Weather App

21 March 14:39

Volume

0

□ Color Tracking Enabled

Temperature units

● C ○ F ○ K

[14 marks]

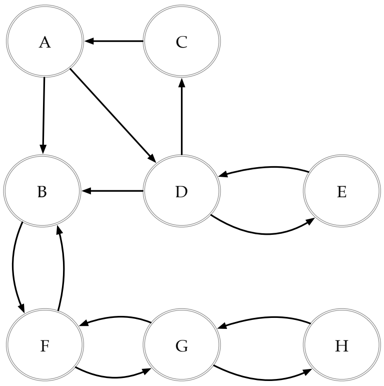
(*b*) Using the outcome of the CW you have conducted above, provide a list of suggestions for re-designing this weather app. [3 marks]

(*c*) Considering that user-centred design is iterative, how would you go about gathering another round of data from the target user group using your findings from the CW? Explain which data gathering techniques you would use and why. [3 marks]

SECTION D

7 Machine Learning and Real-world Data

Consider the directed graph shown in the figure below, which expresses cooperation amongst individuals (A, B, . . . , H) in a fishing village. The meaning of an edge from X to Y is that X has asked Y for advice or help during fishing at least once.



(*a*) Consider the betweenness centrality of each individual in this network, which is listed in the following table.

A	B	C	D	E	F	G	H
6	12	2	9	0	12	8	0

(*i*) Give a definition of the betweenness centrality of a node. [1 mark]

(*ii*) Explain intuitively why B and F have the highest betweenness centralities and why E and H have betweenness centralities of 0. [2 marks]

(*b*) We now look at what happens if the network is converted into an undirected network.

(*i*) What is the diameter of this network and why? Your question should include a definition of diameter. [2 marks]

(*ii*) Do the betweenness centralities of nodes A and C change, and why? Explain in terms of affected paths. [3 marks]

(*iii*) Consider the general case of two near-identical graphs S and T, where S is a directed graph and T is the undirected version of S, i.e., every edge (u, v) in S is replaced by an undirected edge (u, v) in T. Which of the following statements are true about the betweenness centrality of any pair of nodes X_S and X_T, which are in identical relative position in the graphs? Justify your answer or provide a counter example.

[continued . . .]

(A) The betweenness centrality of X_S is always at least that of X_T. [2 marks]

(B) The betweenness centrality of X_S is always equal to that of X_T. [1 mark]

(C) The betweenness centrality of X_S is always at most that of X_T. [2 marks]

(*c*) In directed graphs, the in-degree of a node v is defined as the number of incoming edges (u, v), whereas the node's out-degree is defined as the number of outgoing edges (v, u).

(*i*) What does high in-degree and out-degree mean in the context of the fishing collaboration? [2 marks]

(*ii*) Directed graphs are called "strongly connected" if there exists a path from every node to every other node. Is the graph in Figure 1 strongly connected? Justify your answer. [2 marks]

(*iii*) What is the relation between strong connectedness of a directed graph and its nodes' in- and out-degrees? [3 marks]

8 Machine Learning and Real-world Data

You are a 22nd century historian researching the "FEE" (First Epidemic Era) of 2019–2025, for which records are patchy. You research which government policy was in place in any given week during this historic phase. Policies, in order of severity, are: No restrictions, Tier 1, Tier 2, Tier 3, and Lockdown.

(*a*) From other historic sources, you know the following about sequences of policy levels: if you are in a given policy level, there is a 40% chance you will stay there, a 20% chance that you will be upgraded to the next-highest (more severe) level next week, and a 10% chance that you will be downgraded to the next-lowest (less severe) policy level. The background lockdown probability (which applies if nothing more informative is known about lockdown) is 10%. For each observation sequence, there is also a 5% chance of the sequence ending at any point. Transitions to any other policy level beyond those already described are equally likely. Observation sequences begin with each policy level at equal likelihood.

Using the information given above, construct the full transition probability table.

[7 marks]

(*b*) You want to estimate which policy was in place for the first six weeks of 2025, but unfortunately, the only information you have about this is a sequence of Covid case numbers for these six weeks:

$[0 - 99]$, $[0 - 99]$, $[> 200]$, $[> 200]$, $[> 200]$, $[100 - 199]$.

You know that case loads are associated with policy levels as in the Table below. Describe how you can calculate the sequence of most likely policy levels for these 6 weeks, giving numbers for at least three steps of the calculation. Assume that all policies are equally likely in the week preceding the first week. [8 marks]

	No Restriction	Tier 1	Tier 2	Tier 3	Lockdown
0-99 cases	5%	10%	20%	50%	90%
100-199 cases	15%	40%	40%	30%	9%
> 200 cases	80%	50%	20%	20%	1%

(*c*) In which respects is the modelling described above not fully adequate to describe an actual epidemic? [5 marks]

9 Machine Learning and Real-world Data

You work at a social media company, and your task is to detect cyberbullying messages based on the text they contain. You have access to a large number of messages, N, which have been manually labelled as "OK" and "bullying".

(*a*) How can you apply a Naive Bayes classifier to the task and evaluate it? Describe the approach, including how you would estimate the parameters. [3 marks]

(*b*) You decide to use precision and recall instead of accuracy as the evaluation metric for this task. Why does this decision make sense, and how are the metrics calculated? [1 mark]

(*c*) Your colleague wants to hire two more human annotators to re-label your training data. Why might this be a good idea, how would you measure agreement in this task, and do you think this would improve your classifier in any way? [3 marks]

(*d*) Due to repeated media coverage of cyberbullying, your company introduces a new policy stating that as many cyberbullying messages as possible are to be found and deleted, while still making sure the number of non-filtered messages remains high. Some additional manual labour is made available for this change. How does this affect your evaluation strategy, and how can you adapt the classifier to comply better with the new strategy? (Tip: a development corpus could be of use.) [4 marks]

(*e*) You realise that in this particular application, language change might cause the performance of your trained classifier to drop considerably over time. You have some manual labour available to address the problem, but not enough to relabel large amounts of text.

(*i*) Why is language change relevant here, and how might you notice it? [2 marks]

(*ii*) How could you efficiently build a "cyberbullying lexicon" containing words that have been found in recent cyberbully incidents? [3 marks]

(*iii*) How can you use lexicon-based information to make your classifier more robust towards language change? [4 marks]

END OF PAPER