COMPUTER SCIENCE TRIPOS Part IA 75%, Part IB 50% – 2018 – Paper 3

1 Databases (TGG)

(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]