The assignment is to be turned in before Midnight (by 11:59pm) on January 29th, 2015. Late submissions are accepted, but there is a 5 point deduction for each day the assignment is late. You should turn in the solutions to this assignment as a pdf file through the TEACH website. The solutions should be produced using editing software programs, such as LaTeX or Word, otherwise they will not be graded.

1: Relational Model (3 points)

Briefly describe the concepts of order independence and index independence and explain why relational queries are order and index independent.

solution:

The results of queries over a data management systems should be independent of the order of records and the structure of indexes built on the data. Because results of relational queries, e.g. relational algebra or SQL queries, does not depend on the order of tuples in relations and the existence or the structure of indexes over relations, they are order and index independent.

2: Relational Model (1 point)

Briefly explain a disadvantage of relational model.

solution:

In many domains the data is hierarchical. Relational model is not able to naturally represent the hierarchical structure of these data sets. Any other reasonable explanation is OK.

3: Schema Transformation (2 points)

Consider schema S_1 that contains relation $R(A, B, C, D, E)$ and schema S_2 that contains relations $P_1(A, B, C, D)$ and $P_2(A, B, E)$. Explain whether or not S_2 includes S_1. If it is not, explain which integrity constraints should be added to S_2 and/or S_1 so S_2 includes S_1. Also explain which integrity constraints we should add to S_2 and/or S_1 so they become equivalent. You should explain your answers using the concepts of schema inclusion and schema equivalence proposed in *P. Atzeni, et. al., Inclusion and Equivalence Between Relational Database Schemata, TCS, 1982.* You do not need to use a lot of mathematical notations in your answers. It is sufficient to clearly explain the reasons.

solution:

S_2 does not include S_1. We may add the functional dependency $A, B \rightarrow C, D$ and $A, B \rightarrow E$ to both S_1 and S_2 so that S_2 includes S_1. Further, we may add the inclusion dependency that $P1.A = P2.A$ and $P1.B = P2.B$ to S_2 to make S_1 and S_2 equivalent. Other reasonable constraints are acceptable for this question.

4: Schema Transformation (1 point)

Provide an example of horizontal decomposition where the source schema (original scheme) is **not** included in the refined (target) schema and explain why. You do not need to use a lot of

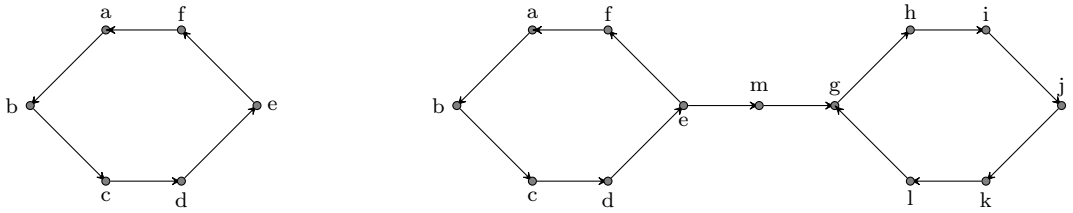
mathematical notations in your answer. It is sufficient to clearly explain the reason.

solution:

Consider source schema S_3 with relation $R(A, B)$ where B can contain any integer value. Assume that the target scheme is S_4 with relations $P_1(A, B)$ and $P_2(A, B)$, such that $P_1.B$ can contain integer values greater than 10 and $P_1.B$ can contain integer values greater than 0. Schema S_3 has some instances that do not map to an instance in S_4, e.g., the instances of S_3 where the value of attribute $R.B$ is negative. Hence, schema S_3 is **not** included in schema S_4. Other reasonable examples are also acceptable.

5: PageRank (3 points)

Assume the following graphs depict parts of the Web, where nodes represent pages and edges show hyper-links. Find out the pages whose PageRank values are greater than zero and their relative PageRank values in both graphs. You do **not** need to perform the fix point computation to determine the PageRank values. Instead, you should guess the PageRank values based on your understanding of the PageRank algorithm and explain why you think they are correct. If it is no possible to make any educated guess for some page(s), you should explain why.



(a) Graph 1 (b) Graph 2

Figure 1: Graphs for Problem 5

solution:

As PageRank algorithm uses teleportation technique, the PageRank value of all pages in both graphs are non-zero. All pages in the first graph have equal PageRank values as they have equal in-degrees and out-degrees. In the second graph, the PageRank values of the nodes g to l are greater than the PageRank values of other nodes in the graph because they trap the random surfer. Because node g have larger in-degree, it is likely to have the largest PageRank value in this group. It is not clear whether pages h to l have equal PageRank values as the links from g might cause h and i to have higher PageRank values than other in this cycle. Similarly, because of the connection from e to g, it is not clear if the nodes d, c, b, a, and f have equal PageRank values.