FACULTE INFORMATIQUE ET COMMUNICATIONS
SCHOOL OF COMPUTER AND COMMUNICATION SCIENCES

Laboratoire de Systèmes d'Information Repartis
Distributed Information Systems Laboratory

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Communication Systems Section
Computer Science Section

Course: Distributed Information Systems

(Karl Aberer)

Test Exam

The following materials are allowed: Script, exercise sheets and solutions, personal notes. Write your name on each solution sheet! The exam consists of 11 sheets including the cover sheet. You can write the answers on the question sheets or use additional sheets. Please number the additional sheets.

NAME: ……..

SECTION: ……..

Please have your student card ready for control!

Each question receives maximal 6 points.

Question 1 Points _____

Question 2 Points _____

Question 3 Points _____

Question 4 Points _____

Total Points _____

GOOD LUCK!

Question 1: Classification Points: ___

The following table consists of training data from a doctor's database:

Database:

Age	**Weight**	**Health**
23	Normal	Healthy
41	Underweight	Sick
66	Normal	Healthy
77	Normal	Sick
30	Overweight	Sick
69	Underweight	Sick
21	Normal	Healthy
55	Normal	Healthy
27	Overweight	Sick
72	Normal	Sick

a) Construct a *Binary Decision Tree* using decision tree induction. "*Health*" is the class label attribute.

b) Estimate the accuracy of the constructed decision tree using the following test set:

Test set:

Age	**Weight**	**Health**
45	Normal	Healthy
48	Underweight	Sick
72	Normal	Sick
26	Overweight	Healthy
61	Overweight	Sick
30	Normal	Healthy

Remark: It is not necessary to compute all the Information Gain values. Compute only those that are necessary for obtaining the correct decisions.

In case your calculator cannot compute Log, here is a table of $\log_2$ values:

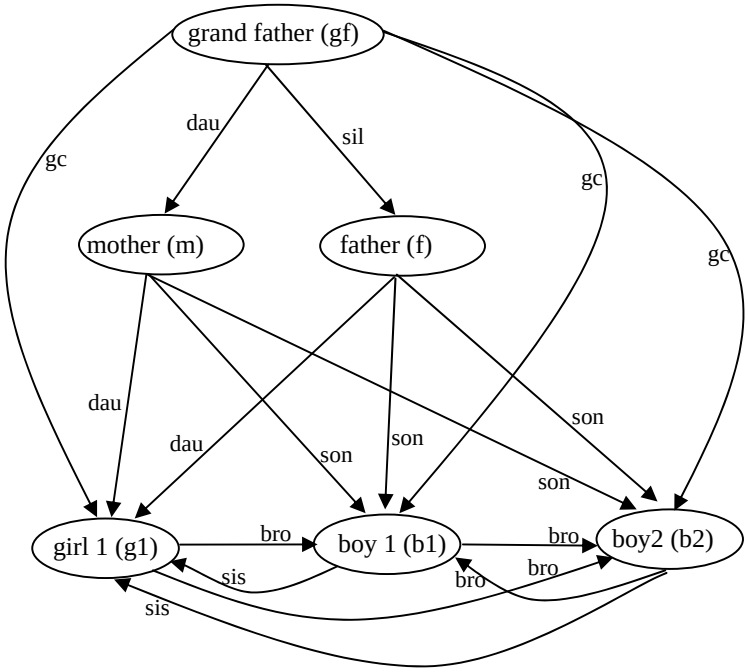
x	**0,025**	**0,05**	**0,075**	**0,1**	**0,125**	**0,15**	**0,175**	**0,2**	**0,225**	**0,25**
$\text{Log}_2(x)$	-5,32193	-4,32193	-3,73697	-3,32193	-3	-2,73697	-2,51457	-2,32193	-2,152	-2
x	**0,275**	**0,3**	**0,325**	**0,35**	**0,375**	**0,4**	**0,425**	**0,45**	**0,475**	**0,5**
$\text{Log}_2(x)$	-1,8625	-1,73697	-1,62149	-1,51457	-1,41504	-1,32193	-1,23447	-1,152	-1,074	-1
x	**0,525**	**0,55**	**0,575**	**0,6**	**0,625**	**0,65**	**0,675**	**0,7**	**0,725**	**0,75**
$\text{Log}_2(x)$	-0,92961	-0,8625	-0,79837	-0,73697	-0,67807	-0,62149	-0,56704	-0,51457	-0,46395	-0,41504
x	**0,775**	**0,8**	**0,825**	**0,85**	**0,875**	**0,9**	**0,925**	**0,95**	**0,975**	**1**
$\text{Log}_2(x)$	-0,36773	-0,32193	-0,27753	-0,23447	-0,19265	-0,152	-0,11247	-0,074	-0,03653	0

Question 2: Semi-structured Data Points: ___

a) Give a short explanation for the following statement: “A data guide is deterministic.”

Given the following data graph:

grandchild (gc)
daughter (dau)
son-in-law (sil)
son (son)
brother (bro)
sister (sis)



b) Construct a data guide for this data graph:

c) Which nodes in the data graph are equivalent from the viewpoint of query processing?

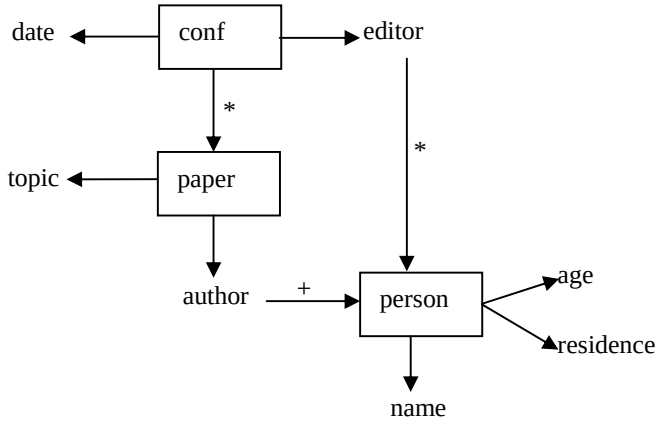
d) Construct a non-deterministic schema graph for the data graph using language equivalence.

e) Assume the non-deterministic schema graph is used to implement an index structure for efficiently processing queries against the data graph. Give the hash table for node gf used for constructing this index structure:

f) Given the following query: /gc/sis
How exactly is this query processed when using the index structure based on the non-deterministic schema graph?

Question 3: XML Storage Points: ___

Given the following DTD graph with the top nodes “conf”, “paper”, and “person”:



Also given is this document:

```
<conf date=”23.02.05”>
  <editor>
    <person name=”Martin” age=65 residence=” Switzerland”>
  </editor>

  <paper topic=”p2p”>
    <author>
      <person name=”Natasha” age=24 residence=”Russia”>
      <person name=”Gilles” age=34 residence=”Switzerland”>
    </author>
  </paper>

  <paper topic=”IR”>
    <author>
      <person name=”Stephanie” age=27 residence=”Germany”>
      <person name=”Jean” age=50 residence=”Switzerland”>
    </author>
  </paper>

  <paper topic=”p2p”>
    <author>
      <person name=”James” age=30 residence=”USA”>
      <person name=”Peter” age=27 residence=”UK”>
      <person name=”Li” age=25 residence=”China”>
    </author>
  </paper>

</conf>
```

When mapping the XML document to a relational representation, the table for top node "person" is:

person						
id	**person_name**	**age**	**residence**	**root_type**	**parent_type**	**parent_element**
pe1	Martin	65	Switzerland	conf	conf	c1
pe2	Natasha	24	Russia	conf	paper	pa1
pe3	Gilles	34	Switzerland	conf	paper	pa1
pe4	Stephanie	27	Germany	conf	paper	pa2
pe5	Jean	50	Switzerland	conf	paper	pa2
pe6	James	30	USA	conf	paper	pa3
pe7	Peter	27	UK	conf	paper	pa3
pe8	Li	25	China	conf	paper	pa3

a) What are the (populated) tables for top nodes "conf" and "paper"?

b) Consider the following XPath queries on the documents:

```
//person[@residence = "Switzerland"]/name

//person[@age < 28]/name

//person[@name = "James"]/residence
```

Formulate corresponding SQL queries on the relational representation of the XML document.

c) Assume you want to produce a horizontal fragmentation of table “person” for the given set of queries. List the simple predicates in these queries:

d) Use the MinFrag algorithm to get the minimal complete set of predicates. Give the resulting horizontal fragments of table “person”.

d) Can you also give a derived horizontal fragmentation for tables “conf” and “paper”?

Question 4: Small-world graphs Points: ___

a) Given a lattice of $(2n+1)^2$ peers in a $(2n+1) \times (2n+1)$ square. Each peer is identified by the coordinates (i, j): $i \in \{1, 2, \ldots 2n+1\}$, $j \in \{1, 2, \ldots 2n+1\}$
The distance between two peers is given by the L_∞ norm, i.e.

$$d((i, j), (k, l)) = max(|k- i|,| l-j|).$$

Assume that a peer has exactly one long-range link, which is chosen according to the optimal construction of small-world graphs as given by Kleinberg. What is the probability that the peer in the centre of the Grid will reach a peer at one of the four corners of the square in exactly one hop?