CS345 Midterm Examination

Wednesday, May 14, 2003, 9:30 – 11:30AM

Directions

- The exam is *open book*; any written materials may be used.
- Answer all 9 questions on the exam paper itself.
- The total number of points is 120 (i.e., 1 point per minute).
- Do not forget to **sign the pledge** below.

I acknowledge and accept the honor code.

_____

Print your name here: _____

1	2	3
4	5	6
7	8	9

Problem 1: (12 points) Let C_1 and C_2 denote two columns (items) of a matrix that represents market-basket data. Let $C_1 \vee C_2$ denote a column that is the row-wise logical OR of the two columns; i.e., $C_1 \vee C_2$ has a 1 when either C_1 or C_2 or both has a 1, and has 0 otherwise. Similarly, $C_1 \wedge C_2$ denotes the row-wise logical AND of the two columns; i.e., $C_1 \wedge C_2$ has a 1 if and only if both columns have 1. Let $h(C)$ denote the minhash value for column C. That is, $h(C)$ is the smallest i such that the ith row in the chosen (permuted) order of rows has a 1 in column C. For each of the following statements, indicate whether it is **always** true or sometimes false, by circling T or F, respectively.

a) $h(C_1 \vee C_2) = \min\big(h(C_1), h(C_2)\big)$ T F

b) $h(C_1 \wedge C_2) = \max\big(h(C_1), h(C_2)\big)$ T F

c) If $h(C_1) = h(C_2)$, then $h(C_1 \vee C_2) = h(C_1)$ T F

d) If $h(C_1) = h(C_2)$, then $h(C_1 \wedge C_2) = h(C_1)$ T F

Problem 2: (12 points) Let $In(x)$ denote the set of pages with link to page x, and let $Out(x)$ denote the set of pages to which page x links. Let $h(x)$, $a(x)$, and $p(x)$ denote the "hubbiness," authority, and PageRank of page x, respectively. Indicate whether each of the following statements is **always** true (T) or sometimes false (F).

a) If $Out(i) \subseteq Out(j)$, then $h(i) \leq h(j)$. T F

b) If $Out(i) \subseteq Out(j)$, then $p(i) \leq p(j)$. T F

c) If $In(i) \subseteq In(j)$, then $p(i) \leq p(j)$. T F

d) If $In(j) \subseteq In(i)$, then $a(i) \leq a(j)$. T F

Problem 3: (15 points) What are all the stable models for the following propositional-logic program?

```
p1 :- NOT q1
q1 :- NOT p1
p2 :- p1
p2 :- NOT q2
q2 :- NOT p2
```

Problem 4: (15 points) A collection of market-basket data has 100,000 frequent items, and 1,000,000 infrequent items. Each pair of frequent items appears 100 times; each pair consisting of one frequent and one infrequent item appears 10 times, and each pair of infrequent items appears once. Answer each of the following questions. Your answers only have to be correct to within 1%, and for convenience, you may optionally use scientific notation, e.g., 3.14×10^8 instead of 314,000,000.

a) What is the total number of pair occurrences? That is, what is the sum of the counts of all pairs?

---

b) We did not state the support threshold, but the given information lets us put bounds on the support threshold s. What are the tightest upper and lower bounds on s?

---

c) Suppose we apply the PCY algorithm to this data. If the actual support threshold s is 10,000,000 (i.e., 10^7), and pairs in each of the three categories distribute as evenly as possible, what is the smallest number of buckets we can use so that most of the buckets are not frequent?

Problem 5: (16 points) Consider the following rules:

```
p(X) :- int(X) & X ≥ 2 & NOT c(X)
c(X) :- int(X) & p(Y) & divides(X,Y) & X ≠ Y
```

Think of $p(X)$ as meaning "X is a prime" and $c(X)$ as "X is composite." The EDB predicate $int(X)$ says that X is a positive integer, and in practice it will hold a finite set of integers. The EDB predicate $divides(X, Y)$ means that Y evenly divides X.

Suppose that $int = \{1, 2, 3, 4\}$, and $divides$ is the expected relation on these four integers; that is, $divides = \{(1,1), (2,1), (3,1), (4,1), (2,2), (4,2), (3,3), (4,4)\}$. If we instantiate these rules in all possible ways, eliminate rules with a known false subgoal and then eliminate known true subgoals from the remaining rules, we are left with the following:

```
p(2) :- NOT c(2)     c(2) :- p(1)
p(3) :- NOT c(3)     c(3) :- p(1)
p(4) :- NOT c(4)     c(4) :- p(1)
                     c(4) :- p(2)
```

a) Use the alternating-fixedpoint method to compute the well-founded model for this program plus EDB, by filling in the following table and then indicating the truth value (T, F, UNK) of each of the eight ground atoms. The table may have extra space for rounds that need not be computed; you may fill in the table only until you are

sure you have reached convergence.

Round	0	1	2	3	4	Truth Value
$p(1)$						
$p(2)$						
$p(3)$						
$p(4)$						
$c(1)$						
$c(2)$						
$c(3)$						
$c(4)$						

b) In the space below, draw the dependency graph for the instantiated atoms $p(i)$ and $c(i)$ for $1 \leq i \leq 4$.

c) Are the rules with the given EDB locally stratified? _____ If so, tell what the strata are; if not, describe an infinite negative path.

d) Suppose *int* contains the integers from 1 to n, and *divides* contains all those pairs (i, j) such that j divides i and i and j are integers between 1 and n. For what values of n will the rules and EDB be locally stratified? Explain briefly.

Problem 6: (16 points) A view-centric information system has a single view:

```
v(X,Y,Z) :- e(X,Y) & e(Y,Z) & e(X,Z)
```

We wish to answer the following query:

```
q(A,B,C,D) :- e(A,B) & e(B,C) & e(C,D) & e(A,C) & e(B,D) & e(A,D)
```

Notice that in this unusual case, neither the view definition nor the query have any variables that do not appear in the head. That fact may simplify reasoning about the problem. Also observe that the view describes a triangle in a graph, but the edges are directed, and go in the direction from one argument of the head (representing a node) to another that appears to the right, among the arguments of the head. Likewise, the query asks for a complete graph of 4 nodes, again with direction determined by "to the right, among the arguments of the head."

A conjunctive query Q, all of whose subgoals have predicate v, is a *solution* if, after expansion, it is contained in the query. For Q to be a *minimal* solution, any conjunctive query P formed by deleting one or more subgoals from the body of Q must not be a solution; i.e., the expansion of P is not contained in the query. For each of the proposed solutions below, tell whether it is: (i) not a solution, (ii) a solution but not minimal, or (iii) a minimal solution. In each case, explain your reasoning briefly. Suggestions: describe the expansions of the proposed solutions and indicate containment mappings when needed.

a) `q(A,B,C,D) :- v(A,B,C) & v(B,C,D)`

b) `q(A,B,C,D) :- v(A,B,C) & v(B,C,D) & v(A,C,D)`

c) `q(A,B,C,D) :- v(A,B,C) & v(A,E,D) & v(B,F,D) & v(G,C,D)`

d) `q(A,B,C,D) :- v(A,B,C) & v(B,C,D) & v(B,A,D)`

Problem 7: (8 points) A market-basket data set contains 10 items. For a particular sample of the data, the set of all maximal frequent itemsets is precisely the set of all pairs of items. How many itemsets that are subsets of these 10 items will be a part of the negative border (as used in Toivonen's Algorithm)? _____ Explain your answer briefly.

Problem 8: (16 points) Consider the following conjunctive queries with arithmetic:

Q_2: `panic :- a(X,Y) & a(Y,X) & X<Y`
Q_1: `panic :- a(A,B) & a(B,A) & A≠B`

We wish to check whether or not $Q_1 \subseteq Q2$.

a) Rewrite Q_1 and Q_2 as rectified rules.

b) What are all the containment mappings from the uninterpreted subgoals of Q_2 to those of Q_1?

c) Write the statement about arithmetic that must be checked to verify that $Q_1 \subseteq Q_2$.

d) Is the condition of (c) true? _____ Explain briefly.

Problem 9: (10 points) Suppose a Web graph is undirected, i.e. page i points to page j if and only page j points to page i. Are the following statements true or false? Justify your answers briefly.

a) The hubbiness and authority vectors are identical, i.e for each page, its hubbiness is equal to its authority.

b) The matrix M that we use to compute PageRank is symmetric; i.e. $M[i, j] = M[j, i]$ for all i and j.