Query Evaluation on Probabilistic Databases

CSE 544: Wednesday, May 24, 2006

Problem Setting

Tables:

Review

name	rating	p
Monkey Love	good	.5
	fair	.2
	fair	.6
	poor	.9

Movie

title	year	p
Twelve Monkeys	1995	.8
Monkey Love	1997	.4
Monkey Love	1935	.9
Monkey Love Pl	2005	.7

Queries:

A(x,y) :- Review(x,y),
Movie(x,z), z > 1991

Answers:

Top k

title	rating	p
Twelve Monkeys	fair	.53
Monkey Love	good	.42
Monkey Love Pl	fair	.15

Problem: complexity of query evaluation

Two Problems

Fixed schema S, conjunctive query Q(x,y)

Query evaluation problem

Fix answer tuple (a,b)

Given database I, compute Pr(Q(a,b))

Top-k answering problem

Fix k > 0

Given database I, find k answer tuples with highest probabilities

Related Work: DB

- Cavallo&Pitarelli:1987
- Barbara,Garcia-Molina, Porter:1992
- Lakshmanan,Leone,Ross&Subrahmanian:1997
- Fuhr&Roellke:1997
- Dalvi&S:2004
- Widom:2005

Related Work: Logic

- Query reliability [Gradel,Gurevitch,Hirsch'98]
- Degrees of belief [Bacchus,Grove,Halpern,Koller'96]
- Probabilistic Logic [Nielson]
- Probabilistic model checking [Kwiatkowska'02]
- Probabilistic Relational Model [Taskar,Abbeel,Koller'02]

Probabilistic Database

Schema S, Domain D, Set of instances Inst

<u>Definition</u>

Probabilistic database is a probability distribution

$$Pr : Inst \to [0,1], \quad \Sigma_I \, Pr[I] = 1$$

If Pr[I] > 0 then I is called "possible world"

Probabilistic Database

Representation:

- Independent tuples:
  I-database DB over some schema S^{i}

- Independent and disjoint tuples:
  ID-database DB over some schema S^{id}

Semantics:

- DB “means” probability distribution Pr over schema S

Independent Events

- A tuple is in the database with probability p
- Any two tuples are independent events

Representation

I-Databases

Reviews^i(M,S,p)

Movie	Score	P
m42	good	p_1
m99	good	p_2
m76	poor	p_3

Reviews(M,S)

Mov	Scor

Mov	Scor
m42	good

Mov	Scor
m99	good

Mov	Scor
m42	good
m99	good

Mov	Scor
m76	poor

Mov	Scor
m42	good
m76	poor

Mov	Scor
m42	1995
m76	poor

Mov	Scor
m42	good
m99	good
m76	poor

$Pr[I_1]=(1-p_1)*(1-p_2)*(1-p_3)$

$Pr[I_4]=p_1*p_2*(1-p_3)$

$Pr[I_8]=p_1*p_2*p_3$

$Pr[I_1] + Pr[I_2] + ... + Pr[I_8] = 1$

Possible worlds semantics

Disjoint Events

Needed in

- Many-to-1 matchings
- Possible values for attributes [Barbara'92]

Name	Age
John	34 (0.3) 43 (0.7)
Mary	25

Name	Age	P
John	34	0.3
John	43	0.7
Mary	25	1.0

ID-Databases

Activities^{id}

Time^d	Activity	P
t	walk	p_1
t	run	p_2
t+1	walk	p_3

Activities

Time	Act

$Pr[I_1]=$ $(1-p_1-p_2)*(1-p_3)$

Time	Act
t	walk

Time	Act
t	run

$Pr[I_3]=$ $p_2*(1-p_3)$

Time	Act
t+1	walk

Time	Act
t	walk
t+1	walk

$Pr[I_5]=$ p_1*p_3

Time	Act
t	run
t+1	walk

$$Pr[I_1] + Pr[I_2] + ... + Pr[I_6] = 1$$

ID subsumes I

Reviews^{id}

Movie^d	Score^d	P
m42	good	p_1
m99	good	p_2
m76	poor	p_3

=

Reviews^i

Movie	Score	P
m42	good	p_1
m99	good	p_2
m76	poor	p_3

Note:

Reviews^{id}

Movie	Score	P
m42	good	p_1
m99	good	p_2
m76	poor	p_3

means <u>all</u> tuples are disjoint

Queries

Syntax: conjunctive queries over schema S

```
Q(y) :- Movie(x,y), Review(x,z), z >= 3
```

$Movie^i$

id	year	p
m42	1995	0.95
m99	2002	0.65
m76	2002	0.1
m05	2005	0.7

$Review^i$

mid	rating	p
m42	4	0.7
m42	5	0.45
m99	5	0.82
m99	4	0.68
m05	5	0.79

Two Query Semantics

Possible answer sets

- Given set A: $\Pr[\{t \mid I \models Q(t)\} = A]$
- Used for views

This talk

Possible tuples

- Given tuple t: $\Pr[I \models Q(t)]$
- Used for query evaluation and top-k

Query Semantics

id	year
m42	2004
m99	1901
m76	1902

p_1

id	year
m99	1935
m05	1903

p_2

id	year
m76	1995
m99	1935
m05	2004

p_3

id	year
m87	1934
m44	1904

p_4

Q(y) :- Movie(x,y), Review(x,z)

year	p
1935	$p_2 + p_3 =$ 0.6
2004	$p_1 + p_3 =$ 0.5
1995	$p_3 =$ 0.2
. . .	. . .

top k

Tuple probabilities

Summary on Data Model

- Data Model:
  Semantics = possible worlds
  Syntax = I-databases or ID-databases
- Queries:
  Syntax = unchanged (conjunctive queries)
  Semantics = tuple probabilities

Problem Definition

Fix schema S, query Q, answer tuple t

Problem: given I/ID-database DB, compute Pr[I ⊨ Q(t)]

notation: Pr[Q(t)]

Conventions:

For upper bounds (P or #P): probabilities are rationals

For lower bounds (#P): probabilities are 1/2

Query Evaluation on I-Databases

Outline

- Intuition
- Extensional plans: PTIME case
- Hard queries: #P-complete case
- Dichotomy Theorem

Q(y) :- Movie(x,y), Review(x,z)

Intuition

Movie[i]

id	year	p
m42	1995	p_1
m99	2002	p_2
m76	2002	p_3
m05	2005	p_4

Review[i]

mid	rate	p
m42	4	q_1
m42	2	q_2
m42	3	q_3
m99	1	q_4
m99	3	q_5
m76	5	q_6

Answer

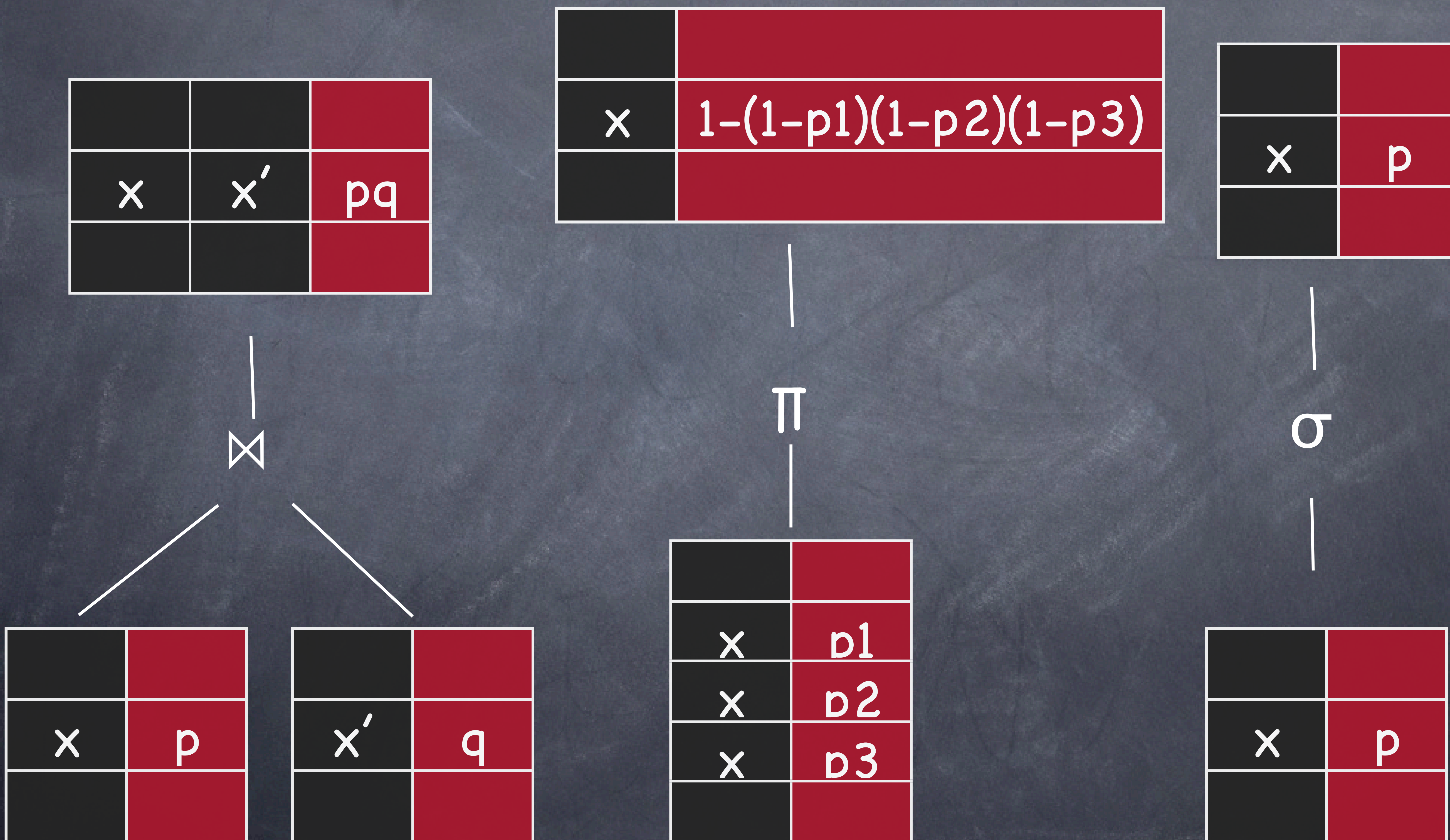
Year	p
1995	$p_1 \times (1 - (1 - q_1) \times (1 - q_2) \times (1 - q_3))$
2002	$1 - (1 - p_2 \times (1 - (1 - q_4) \times (1 - q_5))) \times (1 - p_3 \times q_6)$

I-Extensional Plans

[Barbara92,Lakshmanan97]

- Add p
  Join $\bowtie$ $p = p_1 * p_2$
  Projection Π $p = 1-(1-p_1)(1-p_2)\ldots(1-p_n)$
  Selection σ $p = p$
- Note: data complexity is PTIME

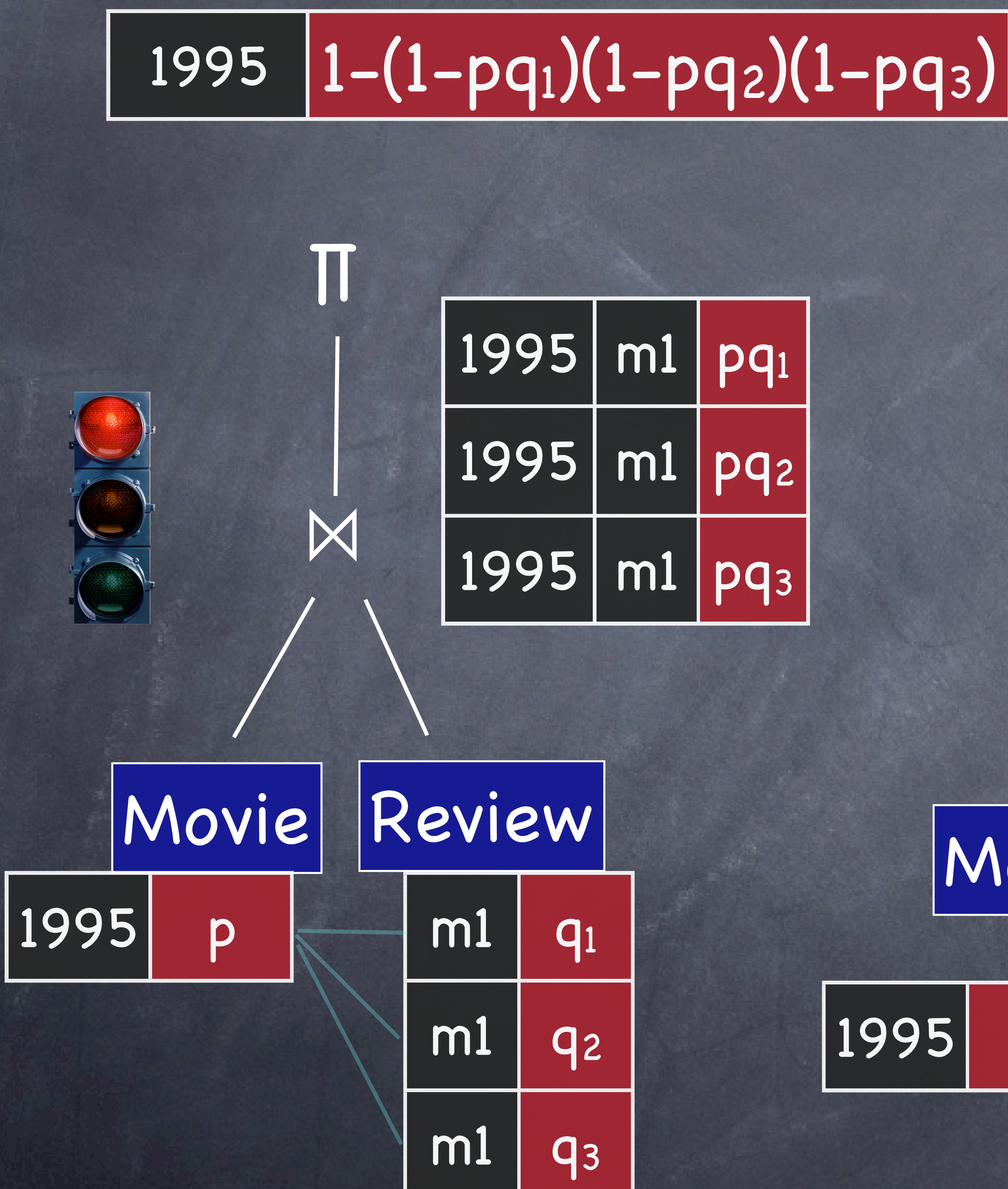
Extensional Query Plans



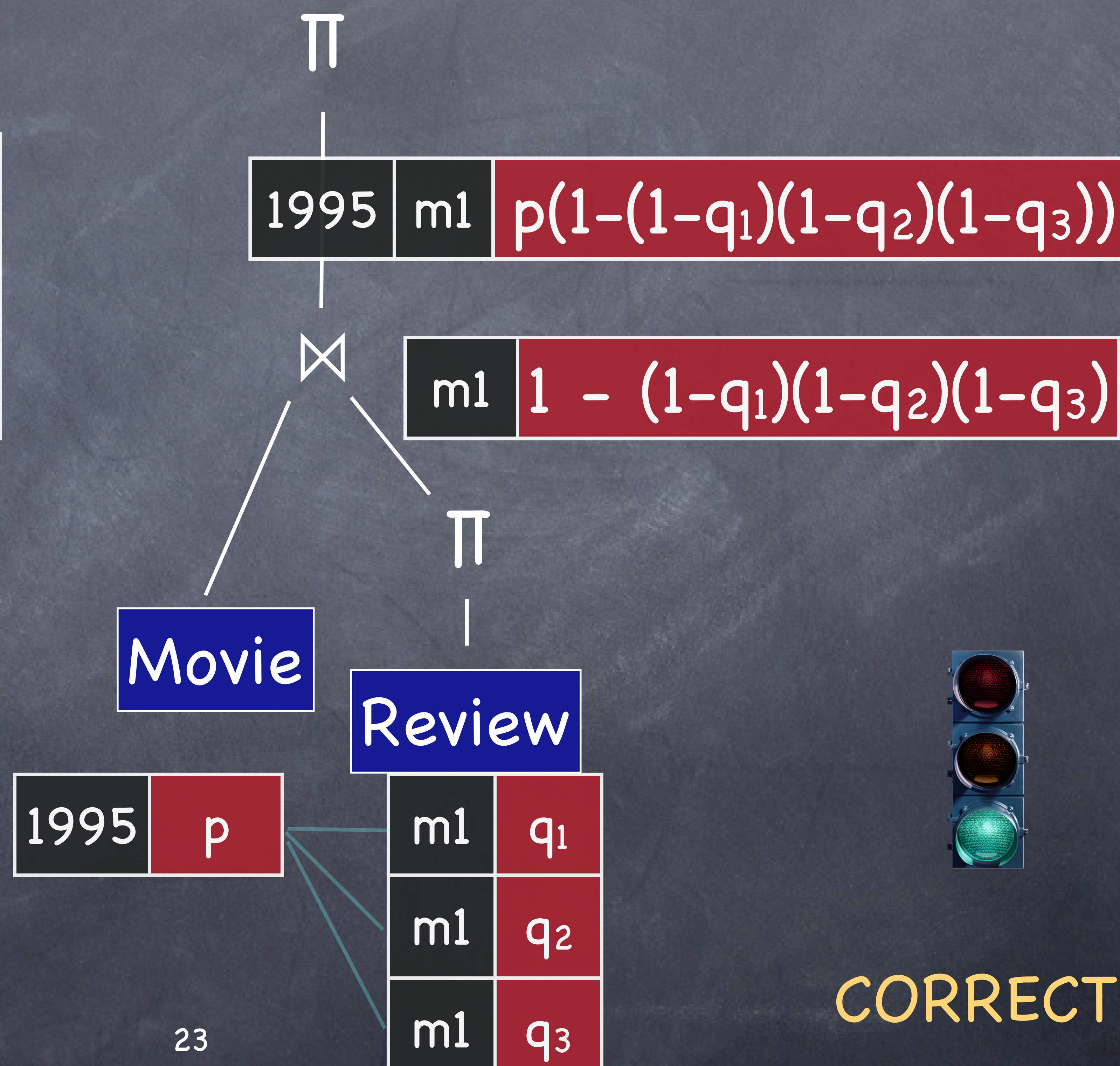
Extensional Query Plans

- Each tuple t has a probability t.P
- Algebra operators compute t.P
- Data complexity: PTIME

$Q(y) \text{ :- } Movie(x,y), Review(x,z)$

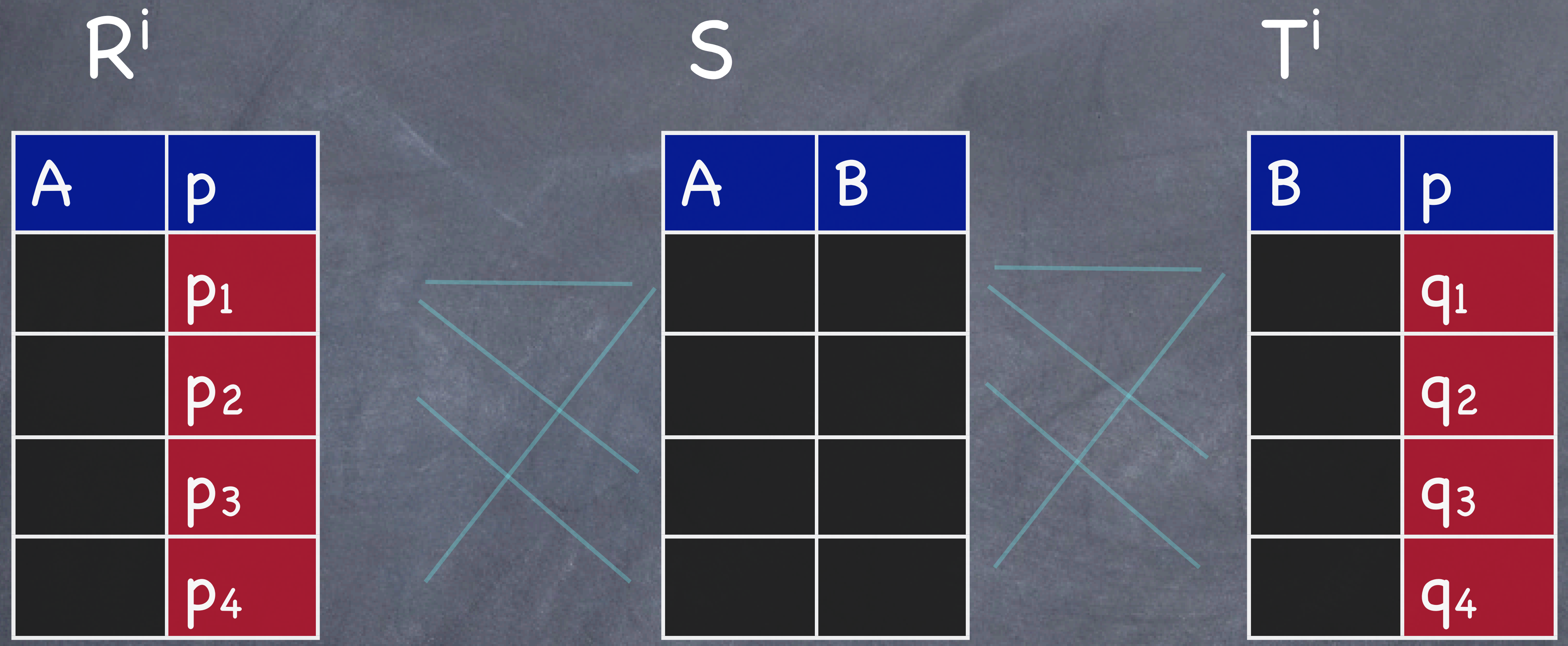


INCORRECT!



CORRECT

#P-Complete Queries



Q_{bad} :- $R^i(x)$, $S(x,y)$, $T^i(y)$

Theorem: Data complexity is #P-complete

Proof:

> **Theorem** [Provan&Ball83] Counting the number of satisfying assignments for bipartite DNF is #P-complete

Reduction:

$x_2y_3 \vee x_1y_2 \vee x_4y_3 \vee x_3y_1$

R^i

A	p
x_1	1/2
x_2	1/2
x_3	1/2
x_4	1/2

S

A	B
x_2	y_3
x_1	y_2
x_4	y_3
x_3	y_1

T^i

B	p
y_1	1/2
y_2	1/2
y_3	1/2

Q_{bad} :- $R^i(x)$, $S(x,y)$, $T^i(y)$

I-Dichotomy

Q = boolean conjunctive query

Definition 1. For each variable x:
goals(x) = set of goals that contain x

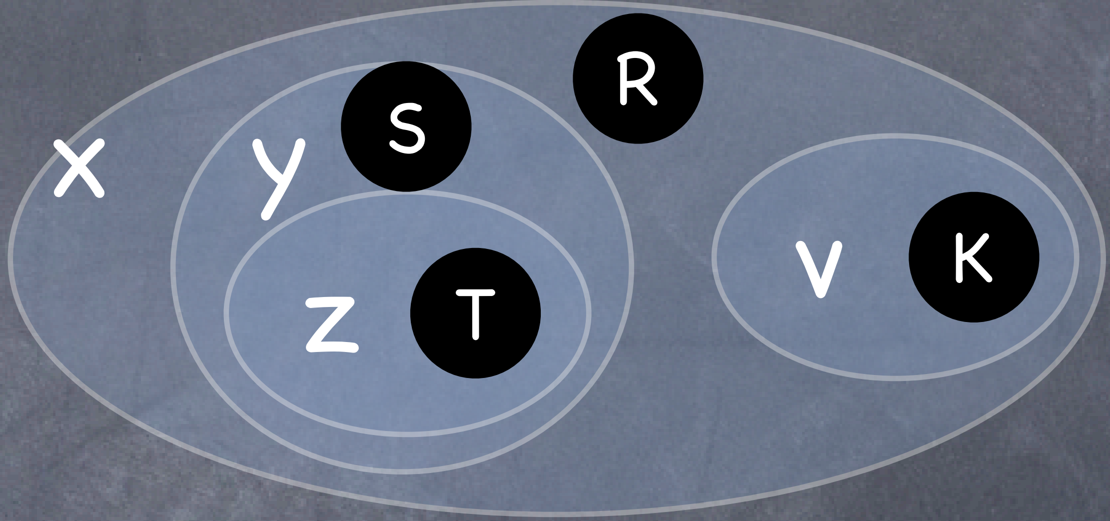
Definition 2. Q is hierarchical if forall x, y:

(a) $goals(x) \cap goals(y) = \emptyset$, or

(b) $goals(x) \subseteq goals(y)$, or

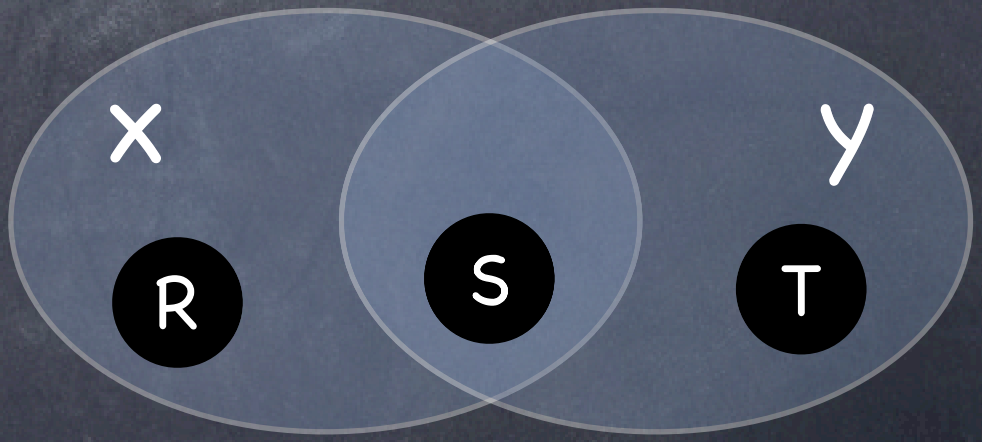
(c) $goals(y) \subseteq goals(x)$

Q :- R(x),S(x,y),T(x,y,z),K(x,v)



"hierarchical"

Q :- R(x), S(x,y), T(y)



"non-hierarchical"

[Dalvi&S.'04]

I-Dichotomy

Schema $S^i = \{R_1^i, R_2^i, \ldots, R_m^i\}$

Theorem Let Q = conjunctive query w/o self-joins.
Then one of the following holds:

Q is in PTIME
Q has a correct extensional plan
Q is hierarchical

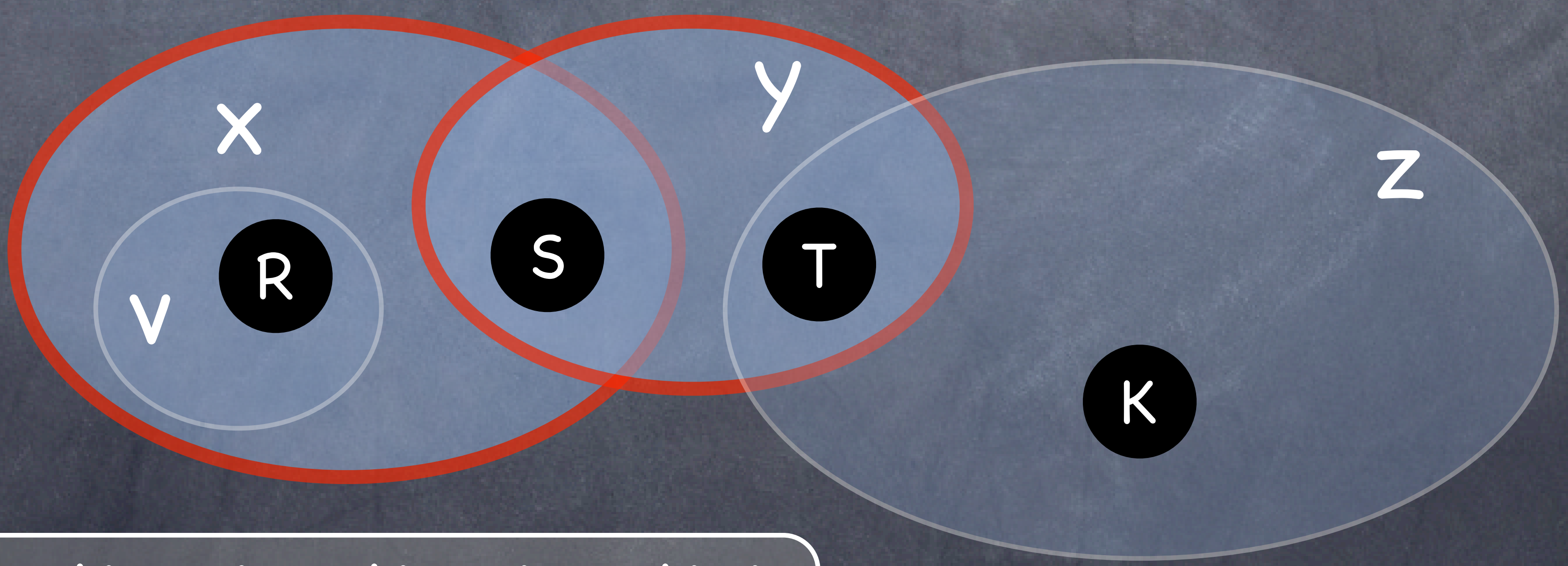
or:

Q is #P-complete
Q has subgoals R(x,...),S(x,y,...),T(y,...)

Proof

Lemma 1.
If Q is non-hierarchical, then #P-complete

Proof:



Q :- $R^i(v,\underline{x})$, $S^i(\underline{x},\underline{y})$, $T^i(\underline{y},z)$, $K^i(z)$

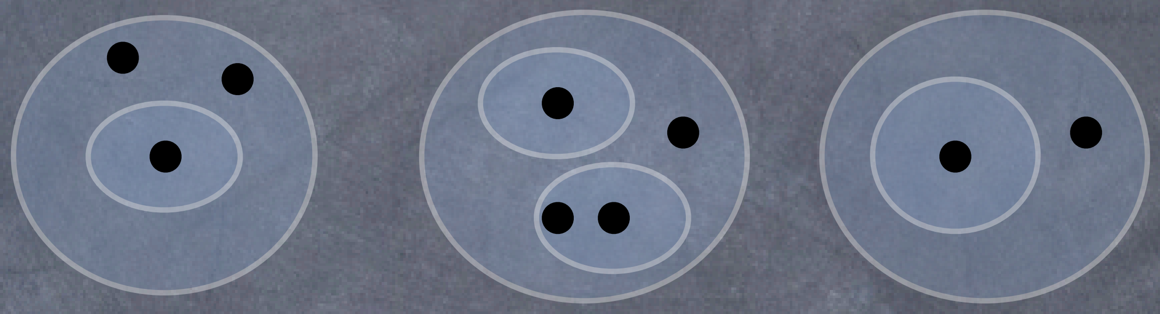
rest is like for Q_{bad}

Proof

Lemma 2. If Q is hierarchical, then PTIME

Proof:

Case 1: has no root



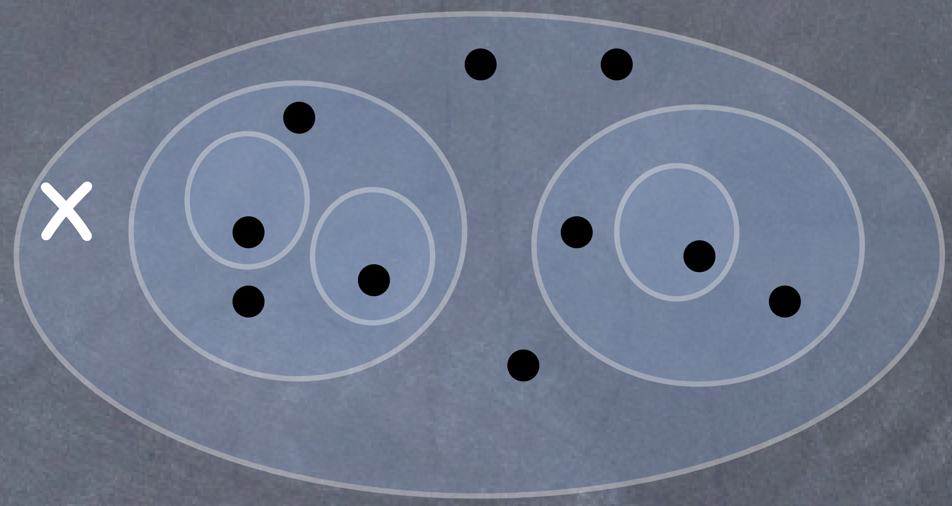
$$Pr(Q) = Pr(Q_1)\ Pr(Q_2)\ Pr(Q_3)$$

This is extensional join ⋈

Proof

Case 2: has root x

Dom={a_1, a_2, . . ., a_n}



Pr(Q) =

$$1 - (1\text{-}Pr(Q(a_1/x))(1\text{-}Pr(Q(a_2/x))\ldots(1\text{-}Pr(Q(a_n/x)))$$

This is an extensional projection: Π

QED

Query Evaluation on ID-Databases

- ID-extensional plans
- #P-complete queries
- Dichotomoy Theorem

Extensional Plans for ID-DBs

- Only difference: two kinds of projections:
  independent $1-(1-p_1)...(1-p_n)$
  disjoint $p_1 + ... + p_n$

#P-Complete Queries

Q_1 :- $R^i(x), S^i(x,y), T^i(y)$

Q_2 :- $R^d(x^d,y), S^d(y^d)$

Q_3 :- $R^d(x^d,y), S^d(z^d,y)$

[Dalvi&S.'04]

I-DB Dichotomy

Schema S^{id} s.t. each table is either R^i or R^{id}

Theorem Let Q = conjunctive query w/o self-joins.
Then one of the following holds:

Q is in PTIME
Q has a correct extensional plan

or:

Q is #P-complete
Q has one of Q_1, Q_2, Q_3 as subqueries

Extensions

Extensions of the dichotomoy theorem exists for:

- Mixed schemas (some relations are deterministic)
- Functional dependencies

Summary on Query Evaluation

Extensional plans: popular, efficient, BUT

- "Equivalent" plans lead to different results
- Some queries admit "correct" plans

Some simple queries: #P-complete complexity

Dichotomy theorem

Future work: remove 'no-self-join' restriction

Conclusions

- Strong motivation from practical applications
  Merge query and search technologies
- Probabilistic DB's are hard !
  Hacks don't work (yet). Need principled approach.

Thank you !

Questions ?