Topics in Probabilistic and Statistical Databases
Lecture 8: Implicit Probabilistic Data

Dan Suciu
University of Washington

Incomplete Databases

- Instance I $\in$ Inst is not known
- Accessed through observations V: views, statistics, constraints, etc.
- User wants to evaluate a query Q

Incomplete Databases

A	B

I_1

A	B

I_2

A	B

I_3

A	B

I_4

A	B

I_5

A	B

I_6

A	B

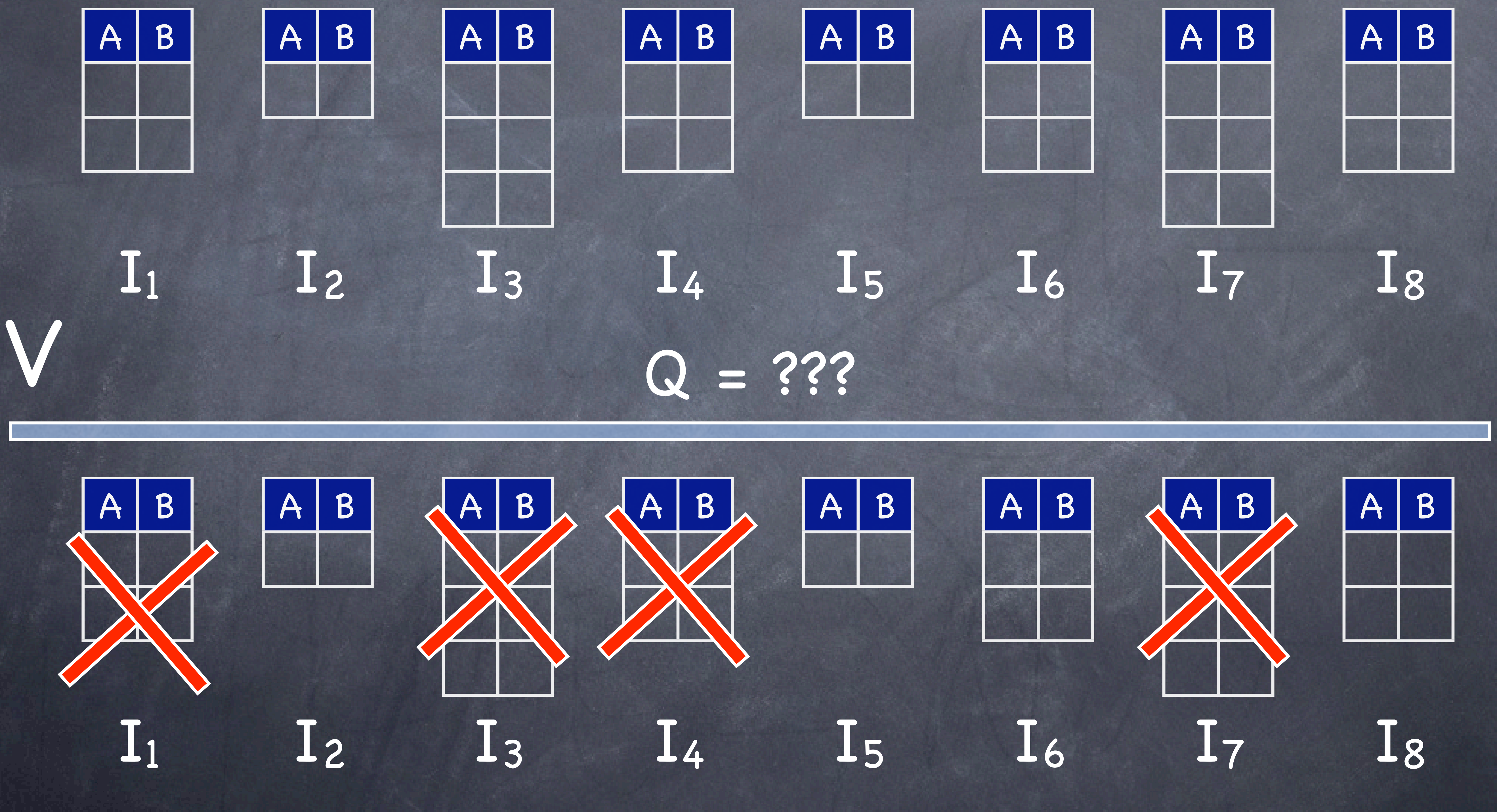
I_7

A	B

I_8

Q = ???

Incomplete Databases



Examples

Information Leakage

- I=private database
- V=published view / anonymized data
- Q=secret query

Examples

Information Leakage

- I=private database
- V=published view / anonymized data
- Q=secret query

Query answering using views

- I=inaccesible database
- V=materialized views
- Q=user query

Examples

Information Leakage

- I=private database
- V=published view / anonymized data
- Q=secret query

Query answering using views

- I=inaccesible database
- V=materialized views
- Q=user query

Size estimation

- I=large database
- V=statistics on I (histograms, samples, etc)
- Q=a count(*) query

Traditional Approach

Classify answers into

- Certain answers
- Possible answers
- Impossible answers

Bayes' Approach

- Assume a prior probability distribution:
  $Pr : Inst \to [0,1], \quad \Sigma_{I \in Inst} Pr(I) = 1$

$Pr(Q)$

- The observations V condition Pr:

$Pr(Q \mid V)$

- Key question: what prior ?

Uniform Prior

Uniform Prior

- Uniform prior $Pr(I_1)=Pr(I_2)= \ldots=Pr(I_m)$

Uniform Prior

- Uniform prior $Pr(I_1)=Pr(I_2)= \ldots =Pr(I_m)$
- Then each tuple t is in the database with probability 1/2 !

Uniform Prior

- Uniform prior $Pr(I_1)=Pr(I_2)= \ldots =Pr(I_m)$
- Then each tuple t is in the database with probability 1/2 !
- This is where Fagin's 0/1 law for FO applies

Uniform Prior

Uniform Prior

- We have seen Fagin's 0/1 Law for FO

Uniform Prior

- We have seen Fagin's 0/1 Law for FO
- Expected database size is huge !

Uniform Prior

- We have seen Fagin's 0/1 Law for FO
- Expected database size is huge !
- Whatever we learn through V is insignificant

Background: Random Graphs

- Relation R of arity k
- Domain size = n
- $N = n^k$ tuples
- Indepedent tuple probability $p \in [0,1]$
- Random graphs:

$$\mu_n : \text{Inst} \rightarrow [0,1], \quad \Sigma_I\ \mu_n(I)=1$$

$$\mu_n(I)=p^{|I|}(1-p)^{N-|I|}$$

Material Random Graphs as Prior

- Relation R of arity k
- Let $\beta > 0$ be a constant
- Tuple probability: $p = \beta/n^k \to 0$
- This defines μ_n
- Note: expected size of the table = β

Notations

- Schema $S = \{R_1, R_2, \ldots, R_m\}$
  arities: $k_1, k_2, \ldots, k_m$
  numbers: $\beta_1, \beta_2, \ldots, \beta_m > 0$
- Defines $\mu_n : Inst \to [0,1]$
- Boolean conjunctive queries Q, V: with constants, joins and ≠

Study:

$$\mu(Q \mid V) = \lim_n \mu_n(Q \mid V) = \lim_n \mu_n(QV) / \mu_n(V)$$

Examples

Convention:

constants: a,b,c,...

variables: x,y,z,...

Q_1 :- R(a,b)

$\mu_n(Q_1) = p = \beta/n^2$

Examples

Convention:

constants: a,b,c,...

variables: x,y,z,...

Q_1 :- R(a,b)

$\mu_n(Q_1) = p = \beta/n^2$

Q_2 :- R(a,-)

Examples

Convention:

constants: a,b,c,...

variables: x,y,z,...

Q_1 :- R(a,b)

$\mu_n(Q_1) = p = \beta/n^2$

Q_2 :- R(a,-)

$\mu_n(Q_2) = 1-(1-p)^n = \beta/n + O(1/n^2)$

Examples

Convention:

constants: a,b,c,...

variables: x,y,z,...

Q_1 :- R(a,b)

$\mu_n(Q_1) = p = \beta/n^2$

Q_2 :- R(a,-)

$\mu_n(Q_2) = 1-(1-p)^n = \beta/n + O(1/n^2)$

Q_3 :- R(a,-), R(-,b)

Examples

Convention:

constants: a,b,c,...

variables: x,y,z,...

Q_1 :- R(a,b)

$\mu_n(Q_1) = p = \beta/n^2$

Q_2 :- R(a,-)

$\mu_n(Q_2) = 1-(1-p)^n = \beta/n + O(1/n^2)$

Q_3 :- R(a,-), R(-,b)

$$\mu_n(Q_3) = 1-(1-p)[1-(1-(1-p)^{n-1})^2] =$$
$$= (\beta + \beta^2)/n^2 + O(1/n^3)$$

Examples

Convention:

constants: a,b,c,...

variables: x,y,z,...

Q_1 :- R(a,b)

$\mu_n(Q_1) = p = \beta/n^2$

Q_2 :- R(a,–)

$\mu_n(Q_2) = 1-(1-p)^n = \beta/n + O(1/n^2)$

Q_3 :- R(a,–), R(–,b)

$$\mu_n(Q_3) = 1-(1-p)[1-(1-(1-p)^{n-1})^2] =$$
$$= (\beta + \beta^2)/n^2 + O(1/n^3)$$

Closed formulas are too complex

Existence Result

Let Q = conjunctive query with constants and ≠

Theorem (1) Given Q, there exists c, e s.t.

$$\mu_n(Q) = c / n^e + O(1/n^{e+1})$$

(2) Given Q, k deciding if e < k is NP-complete

(3) Given Q, computing c is #NP complete

Notation: coeff(Q)=c exp(Q)=c

Case 1: Subgraph Queries

Definition A conjunctive query $Q^{\neq}$ is a subgraph query if it contains all predicates $x \neq v$ for $x \in Vars(Q)$, $v \in Vars(Q) \cup Const(Q)$

$Q_1^{\neq}$:- R(a,b)

$Q_2^{\neq}$:- R(a,x), $x \neq a$

$Q_3^{\neq}$:- R(a,x), R(y,b), $x \neq y$, $x \neq a$, $x \neq b$, $y \neq a$, $y \neq b$

Case 1: Subgraph Queries

Recall the claim: $\mu_n(Q) \approx \text{coeff}(Q) / n^{\exp(Q)}$

Case 1: Subgraph Queries

Recall the claim: $\mu_n(Q) \approx \mathrm{coeff}(Q) / n^{\exp(Q)}$

Theorem Let $Q^{\neq}$ subgraph query*

$\exp(Q) = A(Q) - V(Q)$

$\mathrm{coeff}(Q) = \Pi_{g \in goals(Q)} \beta(g)/\mathrm{Aut}(Q)$

$A(Q) = \Sigma_{g \in goals(Q)} \mathrm{arity}(g)$

$V(Q) = |\mathrm{Vars}(Q)|$

$\mathrm{Aut}(Q) = \#\mathrm{automorphisms}$

*without trivial subgoals

$\exp(Q) = A(Q) - V(Q)$
$\text{coeff}(Q) = \Pi\ \beta(g)$

Examples

$Q_1^{\neq}$:- R(a,b)

$A(Q_1) = 2,\ V(Q_1) = 0$
$\mu_n(Q_1) \approx \beta/n^2$

$Q_2^{\neq}$:- R(a,-)

$A(Q_2) = 2,\ V(Q_2) = 1$
$\mu_n(Q_2) \approx \beta/n$

$Q_3^{\neq}$:- R(a,-), R(-,b)

$A(Q_3) = 4,\ V(Q_3) = 2$
$\mu_n(Q_3) \approx \beta^2/n^2$

Case 2: Conjunctive Queries

Definition Let Q = conjunctive query

UQ(Q) = {S | S = h(Q), h = homomorphism}

$UQ_0(Q)$ = {S | S in UQ(Q), $exp(S^{\neq})$ is min}

Case 2: Conjunctive Queries

Definition Let Q = conjunctive query

$UQ(Q) = \{S \mid S = h(Q), h = \text{homomorphism}\}$

$UQ_0(Q) = \{S \mid S \text{ in } UQ(Q), \exp(S^{\neq}) \text{ is min}\}$

Theorem Let Q= conjunctive query

$\exp(Q) = \min \{\exp(S^{\neq}) \mid S \text{ in } UQ(Q)\}$

$\text{coeff}(Q) = \Sigma \{\text{coeff}(S^{\neq}) \mid S \text{ in } UQ_0(Q)\}$

Examples

Query Q

$UQ_0(Q)$

Q_3 :- R(a,-), R(-,b)

Q_4 :- R(a,b,-), R(-,b,c)

Q_5 :- R(a,x),R(x,y),R(y,b)

Examples

Query Q

$UQ_0(Q)$

Q_3 :- R(a,-), R(-,b)

R(a,-), R(-,b)

R(a,b)

$\mu_n(Q_3) \approx (\beta+\beta^2)/n^2$

Q_4 :- R(a,b,-), R(-,b,c)

Q_5 :- R(a,x),R(x,y),R(y,b)

Examples

Query Q

$UQ_0(Q)$

Q_3 :- R(a,-), R(-,b)

R(a,-), R(-,b)

R(a,b)

$\mu_n(Q_3) \approx (\beta+\beta^2)/n^2$

Q_4 :- R(a,b,-), R(-,b,c)

R(a,b,c)

$\mu_n(Q_4) \approx \beta/n^3$

Q_5 :- R(a,x),R(x,y),R(y,b)

Examples

Query Q	$UQ_0(Q)$	
Q_3 :- R(a,-), R(-,b)	R(a,-), R(-,b) R(a,b)	$\mu_n(Q_3) \approx (\beta+\beta^2)/n^2$
Q_4 :- R(a,b,-), R(-,b,c)	R(a,b,c)	$\mu_n(Q_4) \approx \beta/n^3$
Q_5 :- R(a,x),R(x,y),R(y,b)	R(a,x),R(x,y),R(y,b) R(a,a),R(a,b) R(a,b),R(b,b)	$\mu_n(Q_5) \approx (\beta^3+2\beta^2)/n^4$

Recall: $\mu_n(Q) \approx \text{coeff}(Q) / n^{\exp(Q)}$

In practice: Exp(Q) > 0

- When exp(Q)=0 then A(Q) = V(Q)
- Examples: Q:-R(x,y)   Q':-R(x,y),S(u,v,w)
- Call them "trivial" queries

Consider only non-trivial queries

Recall: $\mu_n(Q) \approx \text{coeff}(Q) / n^{\exp(Q)}$

In practice: Exp(Q) > 0

- When exp(Q)=0 then A(Q) = V(Q)
- Examples: Q:-R(x,y) Q':-R(x,y),S(u,v,w)
- Call them "trivial" queries

Consider only non-trivial queries

Proposition For nontrivial Q: $\lim_{n\to\infty} \mu_n(Q) = 0$

Conditional Probability

Given two conjunctive queries Q, V

- $\mu_n(Q \mid V) = \mu_n(QV) / \mu_n(V)$
- $\mu(Q \mid V) = \lim_{n\to\infty} \mu_n(Q \mid V)$

Conditional Probability

Theorem

(1) μ(Q | V) exists and is:

coeff(QV) / coeff(V) when exp(QV)=exp(V)
0 when exp(QV) > exp(V)

(2) Computing μ(Q | V) is #NP-complete

Two Applications

- Information leakage
- Query answering using views

1. Information Leakage

- Have private instance I
- Want to publish view V(I)
- Doest this leak data about a secret Q(I) ?

Example 1

Employee(name, dept, phone)

V :- Employee('Mary', 'Sales', -)

Q :- Employee('Mary', -, 555123)

Does V leak information about Q ?

Example 2

Employee(name,	dept,	phone)

V :- Employee('Mary', 'Sales', -)
V' :- Employee(- , 'Sales', 555123)

Q :- Employee('Mary', -, 555123)

Do V, V' leak information about Q ?

Background: Perfect Security

[Miklau&S]

Pr : Inst → [0,1] tuple-independent distribution
s.t. Pr(I) ≠ 0 forall I

Definition Q, V are perfectly secure if
Pr[Q | V] = Pr[Q]

Perfect Security

[Miklau&S]

Theorem Q, V are perfectly secure for Pr iff they have no common "critical tuples"

Perfect Security

[Miklau&S]

Theorem Q, V are perfectly secure for Pr iff they have no common "critical tuples"

Theorem If Q,V are perfectly secure some Pr, then they are perfectly secure forall Pr

Example 1

Employee(name, dept, phone)

V :- Employee('Mary', 'Sales', -)

Q :- Employee('Mary', -, 555123)

$Pr[Q \mid V] \neq Pr[Q]$ No perfect security !

"Perfect security" ideal for small domains

Perfect Security

- Drawbacks:
- Classifies as "insecure" views considered Ok in practice
- Does not model collusions

Proposition If both (Q,V_1) and (Q,V_2) are perfectly secure then so is (Q, V_1V_2)

Practical Security

Definition Q, V are practically secure if $\mu(Q \mid V) = 0$ ($= \mu(Q)$)

Theorem [Dalvi&S]
Deciding if Q, V are practically secure $\mu(Q \mid V) = 0$ is Θ_2^p complete

Example 1

Employee(name,	dept,	phone)

V :- Employee('Mary', 'Sales', -)

Q :- Employee('Mary', -, 555123)

$\mu_n(V) \approx \beta/n$ $\mu_n(QV) \approx \beta/n^3$ $\mu(Q \mid V) = 0$

"Practical security" ideal for large domains

Example 2

Employee(name, dept, phone)

V :- Employee('Mary', 'Sales', -)
V' :- Employee(- , 'Sales', 555123)
Q :- Employee('Mary', -, 555123)

$\mu_n(VV') \approx \beta/n^3$ $\mu_n(QVV') \approx \beta/n^3$ $\mu(Q \mid V) = 1$

Explains well collusions

2. Query Answering Using Views

[Levy et al.'95]

- Instance I not accessible
- Have access to view V(I)
- Answer query Q(I) by using only V(I), not I
- Standard approach: certain answers
- For boolean queries: $V \Rightarrow Q$

Example 1

Patient(name, height, weight, disease)

V :- Patient('Mary', 1.65m, 45kg, -)

Q :- Employee('Mary', -, 45kg, -)

Q is a certain answer given V

Example 2

Patient(name, height, weight, disease)

V :- Patient('Mary', 1.65m, 45kg, -)
V' :- Patient(- , 1.65m, 45kg, flu)

Q :- Employee('Mary', -, , flu)

Q is NOT a certain answer given V, V'

Almost Certain Query Answer

Definition Q is a almost certain answer given V if $\mu(Q \mid V) = 1$

Theorem. Deciding a.c. answerability is Π_2^p-complete

Example 2

Patient(name, height, weight, disease)

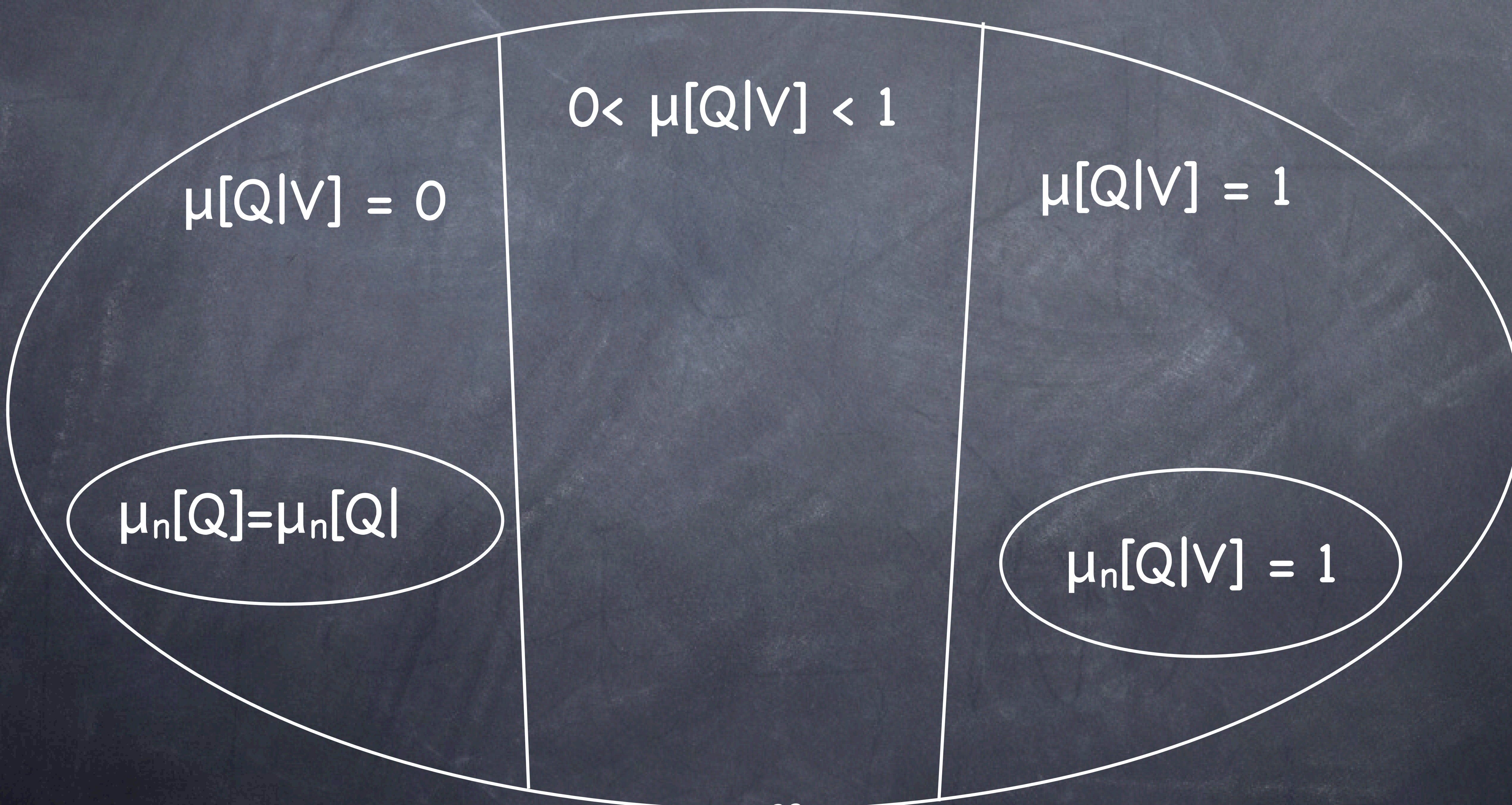
V :- Patient('Mary', 1.65m, 45kg, -)
V' :- Patient(- , 1.65m, 45kg, flu)

Q :- Employee('Mary', -, , flu)

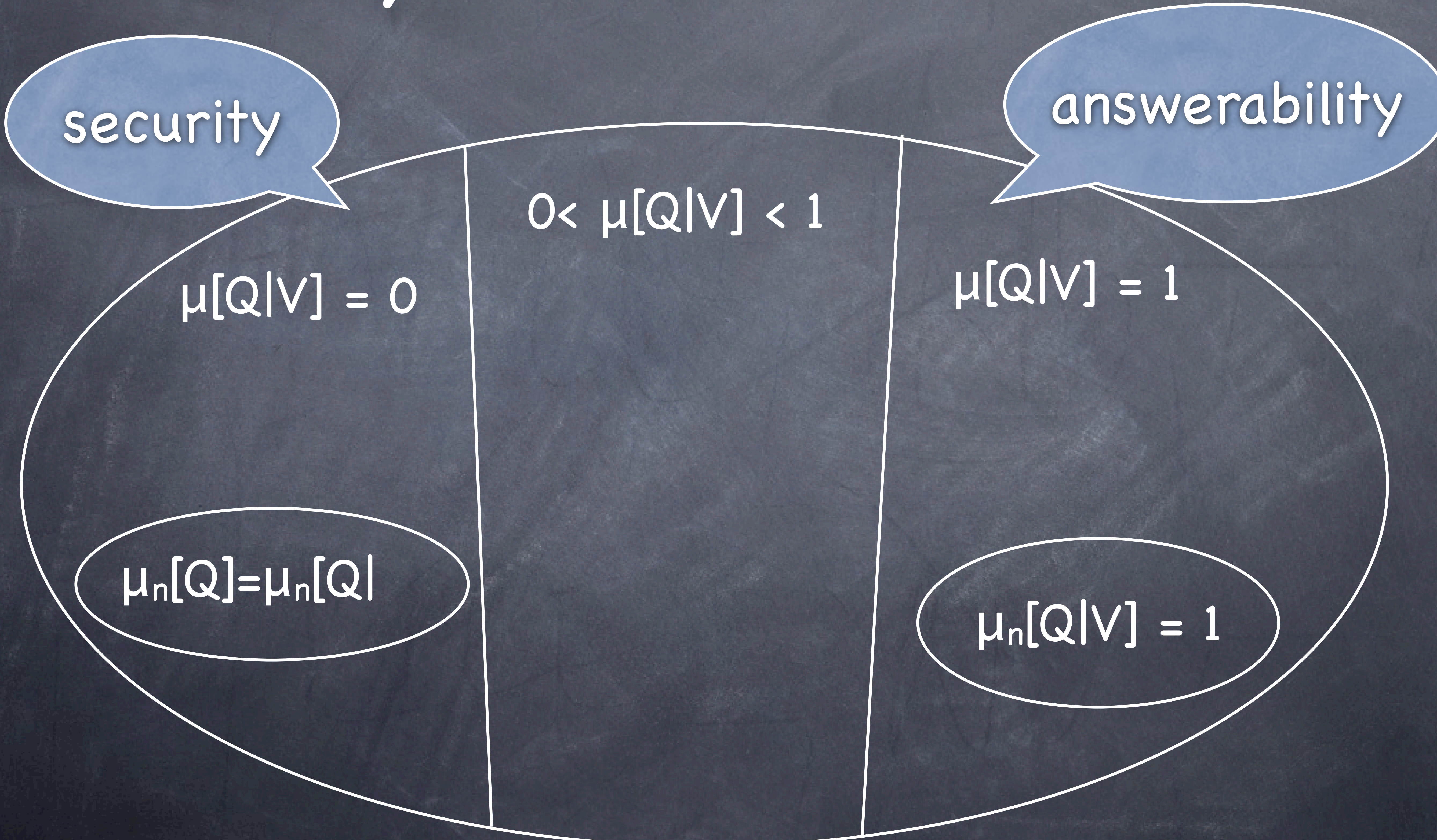
$\mu_n(VV') \approx \beta/n^4$ $\mu_n(QVV') \approx \beta/n^4$ $\mu(Q \mid V) = 1$

Q is an "almost certain" answer given V, V'

Query/View Classification



Query/View Classification



Query/View Classification

security

answerability

$0 < \mu[Q|V] < 1$

Σ^p_2 complete

$\mu[Q|V] = 0$

Θ^p_2 complete

$\mu[Q|V] = 1$

Π^p_2 complete

$\mu_n[Q]=\mu_n[Q|$

Π^p_2 complete

$\mu_n[Q|V] = 1$

NP complete

Summary on Query/View

Application	Probabilities	Complexity	Reference
Perfect security	$\mu_n[Q]=\mu_n[Q \mid V]$	Π^p_2 complete	[Miklau&S]
Practical security	$\mu[Q \mid V] = 0$	Θ^p_2 complete	[Dalvi&S]
Leakage	$0< \mu[Q \mid V] < 1$	Σ^p_2 complete	[Dalvi&S]
Almost certain	$\mu[Q \mid V] = 1$	Π^p_2 complete	[Dalvi&S]
Certain answers	$\mu_n[Q \mid V] = 1$	NP complete	[Duschka]

Advanced Problems

Checking collusions:

Problem Given that $\mu(Q \mid V_1) = 0$, $\mu(Q \mid V_2) = 0$

decide whether $\mu(Q \mid V_1 V_2) = 0$

Advanced Problems

Checking incremental answerability:

Problem Given that $\mu(Q \mid V_1) = 1$

decide whether $\mu(Q \mid V_1 V_2) = 1$

[Dalvi]

Advanced Problems

Three values for $\mu(Q \mid V)$: 0, (0,1), 1

Theorem All 27 combination of values exists:

$\mu(Q \mid V_1)$ $\mu(Q \mid V_2)$ $\mu(Q \mid V_1 V_2)$

[Dalvi]

Advanced Problems

Three values for $\mu(Q \mid V)$:   0,   (0,1),   1

Theorem All 27 combination of values exists:

$\mu(Q \mid V_1)$ $\mu(Q \mid V_2)$ $\mu(Q \mid V_1 V_2)$

Theorem Given known values of

$\mu(Q \mid V_1)$ and $\mu(Q \mid V_2)$

checking $\mu(Q \mid V_1 V_2)$ has same complexity

as checking $\mu(Q \mid V_1 V_2)$.

More Advanced Problems

Relative security

- Some politician approved publishing V, even though Q,V are not practically secure: $\mu(Q \mid V) > 0$
- We want to publish another view, V'
- Problem: is $\mu(Q \mid VV') > \mu(Q \mid V)$?

Theorem Relative security is P.NP-complete

More Advanced Problems

Probabilistic views

- Have explicit probabilities on views: $Pr(V_1) = p_1$, $Pr(V_2) = p_2$, $Pr(V_3) = p_3$
- Use entropy-maximization distribution
- Still possible to compute Pr(Q) assuming views are "non-conflicting": $\mu(V_1 \mid V_2V_3) = \mu(V_2 \mid V_1V_3) = \mu(V_3 \mid V_1V_2) = 0$

[Dalvi&S]