Probabilistic/Uncertain Data Management


Slides based on theSuciu/Dalvi SIGMOD’05 tutorial
What is a Probabilistic Database?

- “An item belongs to the database” is a probabilistic event
  - Tuple-existence uncertainty
  - Attribute-value uncertainty

- “A tuple is an answer to the query” is a probabilistic event

- Can be extended to all data models; we discuss only probabilistic *relational* data
Possible Worlds Semantics

Attribute domains:

```plaintext
int, char(30), varchar(55), datetime
```

# values: $2^{32}$, $2^{120}$, $2^{440}$, $2^{64}$

Relational schema:

```plaintext
Employee(name:varchar(55), dob:datetime, salary:int)
```

# of tuples: $2^{440} \times 2^{64} \times 2^{23}$

Database schema:

```plaintext
Employee(...), Projects(...), Groups(...), WorksFor(...)
```

# of instances: $N$ (= BIG but finite)
The Definition

The set of all possible database instances:

$$\text{INST} = \{I_1, I_2, I_3, \ldots, I_N\}$$

**Definition** A *probabilistic database* $I^p$ is a probability distribution on $\text{INST}$

$$\text{Pr} : \text{INST} \rightarrow [0,1] \quad \text{s.t. } \sum_{i=1,N} \text{Pr}(I_i) = 1$$

**Definition** A *possible world* is $I$ s.t. $\text{Pr}(I) > 0$
Query Semantics

Given a query Q and a probabilistic database $I^p$, what is the meaning of $Q(I^p)$?
Query Semantics

Semantics 1: Possible Answers
A probability distribution on sets of tuples

\[ \forall A. \Pr(Q = A) = \sum_{I \in \text{INST.} \ Q(I) = A} \Pr(I) \]

Semantics 2: Possible Tuples
A probability function on tuples

\[ \forall t. \Pr(t \in Q) = \sum_{I \in \text{INST.} \ t \in Q(I)} \Pr(I) \]
Example: Query Semantics

SELECT DISTINCT x.product
FROM Purchase^p x, Purchase^p y
WHERE x.name = 'John'
and x.product = y.product
and y.name = 'Sue'

Possible answers semantics:

<table>
<thead>
<tr>
<th>Answer set</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo, Camera</td>
<td>1/3</td>
</tr>
<tr>
<td>Gizmo</td>
<td>1/12</td>
</tr>
<tr>
<td>Camera</td>
<td>7/12</td>
</tr>
</tbody>
</table>

Possible tuples semantics:

<table>
<thead>
<tr>
<th>Tuple</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera</td>
<td>11/12</td>
</tr>
<tr>
<td>Gizmo</td>
<td>5/12</td>
</tr>
</tbody>
</table>
Possible Worlds Query Semantics

Possible answers semantics
• Precise
• Can be used to compose queries
• Difficult user interface

Possible tuples semantics
• Less precise, but simple; sufficient for most apps
• Cannot be used to compose queries
• Simple user interface
Possible Worlds Semantics: Summary

*Complete* model; Clean formal semantics for SQL queries

*Not* very useful as a representation or implementation tool

• HUGE number of possible worlds!

Need more effective representation formalisms

• Something that users can understand/explore

• Allow more efficient query execution
  – Avoid “possible worlds explosion”

• *Perhaps giving up completeness*
Representation Formalisms

Problem
Need a good representation formalism

- Will be interpreted as possible worlds
- Several formalisms exist, but no winner
Evaluation of Formalisms

Completeness?
• What possible worlds can it represent?
• What probability distributions on worlds?

Closure?
• Is it closed under evaluation of query operators?
Outline

A complete formalism:
• *Intensional Databases*

Incomplete formalisms:
• Various expressibility/complexity tradeoffs
• Focus on *Explicit Independent Tuples*
Intensional Database

Atomic event ids

\[ e_1, e_2, e_3, \ldots \]

Probabilities:

\[ p_1, p_2, p_3, \ldots \in [0,1] \]

Event expressions: \( \wedge, \vee, \neg \)

\[ e_3 \wedge (e_5 \vee \neg e_2) \]

Intensional probabilistic database \( J \): each tuple \( t \) has an event attribute \( t.E \)
Intensional DB $\Rightarrow$ Possible Worlds

$$J = \begin{array}{|c|c|c|} 
\hline 
\text{Name} & \text{Address} & E \\
\hline 
\text{John} & \text{Seattle} & e_1 \land (e_2 \lor e_3) \\
\text{Sue} & \text{Denver} & (e_1 \land e_2) \lor (e_2 \land e_3) \\
\hline 
\end{array}$$

$$e_1 e_2 e_3 = \begin{array}{cccccccccccccccc} 
000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
\hline 
\end{array}$$

$$\mathcal{I} = \emptyset$$

$$(1-p_1)(1-p_2)(1-p_3) + (1-p_1)(1-p_2)p_3 + (1-p_1)p_2(1-p_3) + p_1(1-p_2)(1-p_3)$$

$$p_1(1-p_2) p_3$$

$$p_1(1-p_2)(1-p_3) + p_1 p_2 p_3$$

$$14$$
Possible Worlds $\Rightarrow$ Intensional DB

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Seattle</td>
</tr>
<tr>
<td>John</td>
<td>Boston</td>
</tr>
<tr>
<td>Sue</td>
<td>Seattle</td>
</tr>
</tbody>
</table>

$p_1$

$E_1 = e_1$
$E_2 = \neg e_1 \land e_2$
$E_3 = \neg e_1 \land \neg e_2 \land e_3$
$E_4 = \neg e_1 \land \neg e_2 \land \neg e_3 \land e_4$

$\text{Pr}(e_1) = p_1$
$\text{Pr}(e_2) = p_2/(1-p_1)$
$\text{Pr}(e_3) = p_3/(1-p_1-p_2)$
$\text{Pr}(e_4) = p_4/(1-p_1-p_2-p_3)$

“Prefix code”

$p_2$

$p_3$

$\text{Pr}(e_1) = p_1$
$\text{Pr}(e_2) = p_2/(1-p_1)$
$\text{Pr}(e_3) = p_3/(1-p_1-p_2)$
$\text{Pr}(e_4) = p_4/(1-p_1-p_2-p_3)$

$p_4$

Intensional DBs are complete
Closure Under Operators

One still needs to compute probability of event expression
Summary on Intensional Databases

Event expression for each tuple
• Possible worlds: any subset
• Probability distribution: any
Complete… but impractical
• Evaluate the probability of long event expressions

Important abstraction: consider restrictions

Related to c-tables  [Imilelinski&Lipski:1984]
Restricted Formalisms

Explicit tuples

• Have a tuple template for every tuple that may appear in a possible world

• Focus on the case of independent tuple events
Explicit Independent Tuples

tuple = independent event

Can be easily extended to capture attribute-value uncertainty
Explicit Independent Tuples

**Tuple independent probabilistic database**

\[
\text{Pr}(I) = \prod_{t \in I} \text{pr}(t) \times \prod_{t \notin I} (1-\text{pr}(t))
\]

\[
\text{INST} = \mathcal{P}(\text{TUP})
\]

\[
N = 2^M
\]

**TUP** = \{t_1, t_2, \ldots, t_M\} = all tuples

\[
\text{pr} : \text{TUP} \rightarrow [0,1]
\]

No restrictions
**Tuple Prob. ⇒ Possible Worlds**

<table>
<thead>
<tr>
<th>Name</th>
<th>City</th>
<th>pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Seattle</td>
<td>p₁ = 0.8</td>
</tr>
<tr>
<td>Sue</td>
<td>Boston</td>
<td>p₂ = 0.6</td>
</tr>
<tr>
<td>Fred</td>
<td>Boston</td>
<td>p₃ = 0.9</td>
</tr>
</tbody>
</table>

\[ J = \]

\[ I^p = \]

\[ \emptyset \]

\[ \begin{array}{ll} 
\text{Name} & \text{City} \\
\hline
\text{John} & \text{Seattle} \\
\text{Sue} & \text{Boston} \\
\text{Fred} & \text{Boston} \\
\end{array} \]

\[ E[\text{size}(I^p)] = 2.3 \text{ tuples} \]

\[ \sum = 1 \]
Tuple-Independent DBs are Incomplete

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Seattle</td>
<td>p₁</td>
</tr>
<tr>
<td>Sue</td>
<td>Seattle</td>
<td>p₂</td>
</tr>
</tbody>
</table>

Very limited – cannot capture correlations across tuples

Not Closed

- Query operators can introduce complex correlations!
 Tuple Prob. $\Rightarrow$ Query Evaluation

<table>
<thead>
<tr>
<th>Name</th>
<th>City</th>
<th>$pr$</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Seattle</td>
<td>$p_1$</td>
</tr>
<tr>
<td>Sue</td>
<td>Boston</td>
<td>$p_2$</td>
</tr>
<tr>
<td>Fred</td>
<td>Boston</td>
<td>$p_3$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Customer</th>
<th>Product</th>
<th>Date</th>
<th>$pr$</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Gizmo</td>
<td>...</td>
<td>$q_1$</td>
</tr>
<tr>
<td>John</td>
<td>Gadget</td>
<td>...</td>
<td>$q_2$</td>
</tr>
<tr>
<td>John</td>
<td>Gadget</td>
<td>...</td>
<td>$q_3$</td>
</tr>
<tr>
<td>Sue</td>
<td>Camera</td>
<td>...</td>
<td>$q_4$</td>
</tr>
<tr>
<td>Sue</td>
<td>Gadget</td>
<td>...</td>
<td>$q_5$</td>
</tr>
<tr>
<td>Sue</td>
<td>Gadget</td>
<td>...</td>
<td>$q_6$</td>
</tr>
<tr>
<td>Fred</td>
<td>Gadget</td>
<td>...</td>
<td>$q_7$</td>
</tr>
</tbody>
</table>

**SELECT DISTINCT** $x$.city  
**FROM** Person $x$, Purchase $y$  
**WHERE** $x$.Name = $y$.Customer and $y$.Product = ‘Gadget’

<table>
<thead>
<tr>
<th>Tuple</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seattle</td>
<td>$p_1(1-(1-q_2)(1-q_3))$</td>
</tr>
<tr>
<td>Boston</td>
<td>$1- (1- p_2(1-(1-q_5)(1-q_6))) \times (1 \cdot p_3 q_7 )$</td>
</tr>
</tbody>
</table>
Application 1: Similarity Predicates

<table>
<thead>
<tr>
<th>Name</th>
<th>City</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Seattle</td>
<td>statistician</td>
</tr>
<tr>
<td>Sue</td>
<td>Boston</td>
<td>musician</td>
</tr>
<tr>
<td>Fred</td>
<td>Boston</td>
<td>physicist</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cust</th>
<th>Product</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Gizmo</td>
<td>dishware</td>
</tr>
<tr>
<td>John</td>
<td>Gadget</td>
<td>instrument</td>
</tr>
<tr>
<td>John</td>
<td>Gadget</td>
<td>instrument</td>
</tr>
<tr>
<td>Sue</td>
<td>Camera</td>
<td>musicware</td>
</tr>
<tr>
<td>Sue</td>
<td>Gadget</td>
<td>microphone</td>
</tr>
<tr>
<td>Sue</td>
<td>Gadget</td>
<td>instrument</td>
</tr>
<tr>
<td>Fred</td>
<td>Gadget</td>
<td>microphone</td>
</tr>
</tbody>
</table>

SELECT DISTINCT x.city
FROM Person x, Purchase y
WHERE x.Name = y.Cust
  and y.Product = 'Gadget'
  and x.profession ~ 'scientist'
  and y.category ~ 'music'
### Application 1: Similarity Predicates

#### Step 1: Evaluate ~ predicates

<table>
<thead>
<tr>
<th>Name</th>
<th>City</th>
<th>Profession</th>
<th>pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Seattle</td>
<td>statistician</td>
<td>p₁=0.8</td>
</tr>
<tr>
<td>Sue</td>
<td>Boston</td>
<td>musician</td>
<td>p₂=0.2</td>
</tr>
<tr>
<td>Fred</td>
<td>Boston</td>
<td>physicist</td>
<td>p₃=0.9</td>
</tr>
</tbody>
</table>

#### Step 2: Evaluate rest of query

<table>
<thead>
<tr>
<th>Cust</th>
<th>Product</th>
<th>Category</th>
<th>pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Gizmo</td>
<td>dishware</td>
<td>q₁=0.2</td>
</tr>
<tr>
<td>John</td>
<td>Gadget</td>
<td>instrument</td>
<td>q₂=0.6</td>
</tr>
<tr>
<td>John</td>
<td>Gadget</td>
<td>instrument</td>
<td>q₃=0.6</td>
</tr>
<tr>
<td>Sue</td>
<td>Camera</td>
<td>musicware</td>
<td>q₄=0.9</td>
</tr>
<tr>
<td>Sue</td>
<td>Gadget</td>
<td>microphone</td>
<td>q₅=0.7</td>
</tr>
<tr>
<td>Sue</td>
<td>Gadget</td>
<td>instrument</td>
<td>q₆=0.6</td>
</tr>
<tr>
<td>Fred</td>
<td>Gadget</td>
<td>microphone</td>
<td>q₇=0.7</td>
</tr>
</tbody>
</table>

**SELECT DISTINCT x.city**
FROM Personᵖ x, Purchaseᵖ y
WHERE x.Name = y.Cust
and y.Product = ‘Gadget’
and x.profession ~ ‘scientist’
and y.category ~ ‘music’

**Tuple** | **Probability**
---|---
Seattle | p₁(1-(1-q₂)(1-q₃))
Boston | 1-(1-p₂(1-(1-q₅)(1-q₆))) × (1-p₃q₇)
Summary on Explicit Independent Tuples

Independent tuples

• Possible worlds: subsets
• Probability distribution: restricted
• Closure: no