Mining Uncertain and Probabilistic Data

Problems, Challenges, Methods, and Applications

Jian Pei¹, Ming Hua¹, Yufei Tao², Xuemin Lin³

¹Simon Fraser University
²The Chinese University of Hong Kong
³The University of New South Wales

Some figures in the slides are borrowed from some papers in the references
Outline

• Uncertainty and uncertain data, where and why?
• Models for uncertain and probabilistic data
• (coffee break)
• OLAP on uncertain and probabilistic data
• Mining uncertain and probabilistic data
• Tools: querying uncertain and probabilistic data
  – Indexing uncertain and probabilistic data
  – Ranking queries and spatial queries
• Summary and discussion
Uncertainty Is (Almost) Everywhere

- Uncertainty is often caused by our limited perception and understanding of reality
  - Limited observation equipment
  - Limited resource to collect, store, transform, analyze, and understand data
- Uncertainty can be inherent in nature
  - How much do you like/dislike McCain and Obama?
Data Collection Using Sensors

• Sensors are often used to collect data
  – Thermal, electromagnetic, mechanical, chemical, optical radiation, acoustic, ...
  – Applications: environment surveillance, security, manufacture systems, ...

• Ideal sensors
  – Ideal sensors are designed to be linear: the output signal of a sensor is linearly proportional to the value of the measured property
  – Sensitivity: the ratio between output signal and measured property
Measurement Errors – Certain

- Sensitivity error: the sensitivity differs from the value specified
- Offset (bias): the output of a sensor at zero input
- Nonlinearity: the sensitivity is not constant over the range of the sensor
Uncertain (Dynamic) Errors

- Dynamic error: deviation caused by a rapid change of the measured property over time
- Drift: the output signal changes slowly independent of the measured property
  - Long term drift: a slow degradation of sensor properties over a long period
- Noise: random deviation of the signal varying in time
- A sensor may to some extent be sensitive to properties (e.g., temperature) other than the one being measured
- Dynamic error due to sampling frequency of digital sensors
Uncertainty in Survey Data

- Social security number: 185 or 785
  - Exclusiveness: SSN should be unique
- Is Smith married?
  - Single or married, but not both

Antova et al. ICDE’07
Uncertainty due to Data Granularity

• Which state is p9 in?
• What is the total repair cost for F150’s in the East?

<table>
<thead>
<tr>
<th></th>
<th>Auto</th>
<th>Loc</th>
<th>Repair</th>
<th>Text</th>
<th>Brake</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>F-150</td>
<td>NY</td>
<td>$200</td>
<td>…</td>
<td>(0.8, 0.2)</td>
</tr>
<tr>
<td>p2</td>
<td>F-150</td>
<td>MA</td>
<td>$250</td>
<td>…</td>
<td>(0.9, 0.1)</td>
</tr>
<tr>
<td>p3</td>
<td>F-150</td>
<td>CA</td>
<td>$150</td>
<td>…</td>
<td>(0.7, 0.3)</td>
</tr>
<tr>
<td>p4</td>
<td>Sierra</td>
<td>TX</td>
<td>$300</td>
<td>…</td>
<td>(0.3, 0.7)</td>
</tr>
<tr>
<td>p5</td>
<td>Camry</td>
<td>TX</td>
<td>$325</td>
<td>…</td>
<td>(0.7, 0.3)</td>
</tr>
<tr>
<td>p6</td>
<td>Camry</td>
<td>TX</td>
<td>$175</td>
<td>…</td>
<td>(0.5, 0.5)</td>
</tr>
<tr>
<td>p7</td>
<td>Civic</td>
<td>TX</td>
<td>$225</td>
<td>…</td>
<td>(0.3, 0.7)</td>
</tr>
<tr>
<td>p8</td>
<td>Civic</td>
<td>TX</td>
<td>$120</td>
<td>…</td>
<td>(0.2, 0.8)</td>
</tr>
<tr>
<td>p9</td>
<td>F150</td>
<td>East</td>
<td>$140</td>
<td>…</td>
<td>(0.5, 0.5)</td>
</tr>
<tr>
<td>p10</td>
<td>Truck</td>
<td>TX</td>
<td>$500</td>
<td>…</td>
<td>(0.9, 0.1)</td>
</tr>
</tbody>
</table>

Burdick et al. VLDB’05
Uncertainty in Data Integration

- Schema 1: (pname, email-addr, permanent-addr, current-addr)
- Schema 2: (name, email, mailing-addr, home-addr, office-addr)
- How to map the two schemas?

<table>
<thead>
<tr>
<th>Possible Mapping</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_1 = {(pname, name), (email-addr, email), (current-addr, mailing-addr), (permanent-addr, home-address)} )</td>
<td>0.5</td>
</tr>
<tr>
<td>( m_2 = {(pname, name), (email-addr, email), (permanent-addr, mailing-addr), (current-addr, home-address)} )</td>
<td>0.4</td>
</tr>
<tr>
<td>( m_3 = {(pname, name), (email-addr, mailing-addr), (current-addr, home-address)} )</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Dong et al. VLDB’07
Ambiguous Entities

• Entity identification is a challenging task
Disguised Missing Data

Information about "State" is missing "Alabama" is used as the disguise

Ah, we should open more branches in Alabama...

Hua and Pei KDD’07
Disguised Missing Data

- Disguised missing data is the missing data entries that are not explicitly represented as such, but instead appear as potentially valid data values
- Disguised missing data also introduces uncertainty
Why Uncertain Data Is Still Useful?

• For a temperature sensor, suppose the difference between the real temperature and the sensed temperature follows normal distribution
• The real temperature can be modeled by a probability density function
• What is the real temperature? Uncertain
• What is the probability that the real temperature is over 50°C? Certain!
Uncertainty and Confidence

• Uncertain data can provide probabilistic answers to aggregate questions
  – How can we estimate the percentage of married voters supporting Obama from survey data?
  – What is the total repair cost for F150’s in the East?
• An answer derived from uncertain data may often be a function on probability or confidence
Reducing Uncertainty

- Removing uncertain entries
  - Removing uncertain attribute values
  - Removing uncertain records
  - Cons: reducing available data

- Generalization
  - Remove attribute city if some entries on the attribute is uncertain
  - Can accurately answer questions at level city or above
  - Still cannot answer questions at level city or below
Being Certain or Uncertain?

• Answering questions on uncertain data in general can be more complicated
  – Probability is a new (and often difficult) dimension

• Simplifying uncertain data to certain data may not use the full potential of data
  – Many details may be lost

• Probabilistic answers on uncertain data are often interesting and useful
Uncertain Data Analysis Framework

- Analytic and data mining tasks
- Probability/confidence-aware queries
- Query answering on uncertain/probabilistic data
- Uncertainty assessment and estimation

Data sources:
- Sensor network
- Survey form
- ...
Uncertain Data Acquisition

- Statistics-based, model-driven approaches are often used
- Misrepresentations of data in sensor networks
  - Impossible to collect all relevant data – potentially infinite
  - Samples are non-uniform in time and space due to non-uniform placement of sensors in space, faulty sensors, high packet loss rates, …
A Model-driven Approach

• Treat each sensor as a variable
  – Hidden variables (e.g., whether a sensor faulty) can be added

• Learn a model (a multivariate probability density function)
  – A machine learning/data mining problem

• Given a query, compute a query plan optimal in communication cost to achieve the specified confidence

Deshpande et al. VLDB’04
Outline

• Uncertainty and uncertain data, where and why?
• Models for uncertain and probabilistic data
• (coffee break)
• OLAP on uncertain and probabilistic data
• Mining uncertain and probabilistic data
• Tools: querying uncertain and probabilistic data
  – Indexing uncertain and probabilistic data
  – Ranking queries and spatial queries
• Summary and discussion
Levels of Uncertainty

- Uncertainty can exist in object/tuple level and attribute level
- Object/tuple level uncertainty
  - An object/tuple takes a probability to appear (existing probability)
- Attribute level uncertainty
  - An attribute of an object/tuple takes a few possible values
### Probabilistic Database Model

#### Speed of cars detected by radar

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Location</th>
<th>Car make</th>
<th>Plate No.</th>
<th>Speed</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>11:45</td>
<td>L1</td>
<td>Honda</td>
<td>130</td>
<td>0.4</td>
</tr>
<tr>
<td>t2</td>
<td>11:50</td>
<td>L2</td>
<td>Toyota</td>
<td>120</td>
<td>0.7</td>
</tr>
<tr>
<td>t3</td>
<td>11:35</td>
<td>L3</td>
<td>Toyota</td>
<td>80</td>
<td>0.3</td>
</tr>
<tr>
<td>t4</td>
<td>12:10</td>
<td>L4</td>
<td>Mazda</td>
<td>90</td>
<td>0.4</td>
</tr>
<tr>
<td>t5</td>
<td>12:25</td>
<td>L5</td>
<td>Mazda</td>
<td>110</td>
<td>0.6</td>
</tr>
<tr>
<td>t6</td>
<td>12:15</td>
<td>L6</td>
<td>Nissan</td>
<td>105</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Generation rules:** $(t2 \oplus t3)$, $(t4 \oplus t5)$

- The values of each tuple are certain
- Each tuple carries an existing/membership probability
- Generation rules: constraints specifying exclusive tuples
Survey Data Example

<table>
<thead>
<tr>
<th>TID</th>
<th>Name</th>
<th>SSN</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>Smith</td>
<td>185</td>
<td>40%</td>
</tr>
<tr>
<td>t2</td>
<td>Smith</td>
<td>785</td>
<td>60%</td>
</tr>
<tr>
<td>t3</td>
<td>Brown</td>
<td>185</td>
<td>50%</td>
</tr>
<tr>
<td>t4</td>
<td>Brown</td>
<td>186</td>
<td>50%</td>
</tr>
</tbody>
</table>

Generation rules:
- t1 $\oplus$ t2,
- t3 $\oplus$ t4,
- t1 $\oplus$ t3

Antova et al. ICDE’07
Uncertain Objects

- An object is uncertain in a few attributes
- Use a sample or a probability density function to capture the distribution on uncertain attributes
Uncertainty of Mobile Objects
Survey Data Example

Object 1: Smith
- Smith.SSN=785, 60%
- Smith.SSN=185, 40%

Object 2: Brown
- Brown.SSN=185, 50%
- Brown.SSN=186, 50%

Constraints:
“Smith.SSN=185” ⊕ “Brown.SSN=185”

Antova et al. ICDE’07
Prob Table vs. Uncertain Objects

• A probabilistic table can be represented as a set of uncertain objects
  – All tuples in a generation rule are modeled as an uncertain object
  – Use NULL instances to make the sum of membership probabilities in one object to 1

• Uncertain objects with discrete instances can be represented using a probabilistic table
  – One record per instance
  – All instances of an object are constrained by one generation rule
  – Uncertain objects with continuous probability density functions cannot be represented using a finite probabilistic table

• More complicated constraints may not be captured in the transformation
## Prob Table vs. Uncertain Objects

### A probabilistic table

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Loc</th>
<th>Car Make</th>
<th>Plate No</th>
<th>Speed</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>11:45</td>
<td>Honda</td>
<td>X-123</td>
<td>130</td>
<td>0.4</td>
</tr>
<tr>
<td>t2</td>
<td>11:50</td>
<td>Toyota</td>
<td>Y-245</td>
<td>120</td>
<td>0.7</td>
</tr>
<tr>
<td>t3</td>
<td>11:35</td>
<td>Toyota</td>
<td>Y-245</td>
<td>80</td>
<td>0.3</td>
</tr>
<tr>
<td>t4</td>
<td>12:10</td>
<td>Mazda</td>
<td>W-541</td>
<td>90</td>
<td>0.4</td>
</tr>
<tr>
<td>t5</td>
<td>12:25</td>
<td>Mazda</td>
<td>W-541</td>
<td>110</td>
<td>0.6</td>
</tr>
<tr>
<td>t6</td>
<td>12:15</td>
<td>Nissan</td>
<td>L-105</td>
<td>105</td>
<td>1.0</td>
</tr>
</tbody>
</table>

*Rules: (t2 ∨ t3), (t4 ∨ t5)*

### A set of uncertain objects

- A tuple
- A generation rule
- An instance
- An uncertain object

![Graph showing speed and time](image-url)

### Notes

- **Time** and **Speed** are measured in minutes and km/h, respectively.
- **Conf** represents the confidence level of the uncertain data.
Possible Worlds

• A possible world
  – a possible snapshot that may be observed
• Probabilistic database model
  – A possible world = a set of tuples
  – At most one tuple per generation rule in a possible world
• Uncertain object model
  – A possible world = a set of instances of uncertain objects
  – At most one instance per object in a possible world
• A possible world carries an existence probability
An Example of Possible Worlds

0.4 = 0.112 + 0.168 + 0.048 + 0.072

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Loc</th>
<th>Car Make</th>
<th>Plate No</th>
<th>Speed</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>L1</td>
<td>Honda</td>
<td>X-123</td>
<td>130</td>
<td>0.4</td>
</tr>
<tr>
<td>t2</td>
<td>L2</td>
<td>Toyota</td>
<td>Y-245</td>
<td>120</td>
<td>0.7</td>
</tr>
<tr>
<td>t3</td>
<td>L3</td>
<td>Toyota</td>
<td>Y-245</td>
<td>80</td>
<td>0.3</td>
</tr>
<tr>
<td>t4</td>
<td>L4</td>
<td>Mazda</td>
<td>W-541</td>
<td>90</td>
<td>0.4</td>
</tr>
<tr>
<td>t5</td>
<td>L5</td>
<td>Mazda</td>
<td>W-541</td>
<td>110</td>
<td>0.6</td>
</tr>
<tr>
<td>t6</td>
<td>L6</td>
<td>Nissan</td>
<td>L-105</td>
<td>105</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Rules: (t2 ⊕ t3), (t4 ⊕ t5)

A probabilistic table

Possible worlds

<table>
<thead>
<tr>
<th>World</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW¹={t1,t2,t6,t4}</td>
<td>0.112</td>
</tr>
<tr>
<td>PW²={t1,t2,t5,t6}</td>
<td>0.168</td>
</tr>
<tr>
<td>PW³={t1,t6,t4,t3}</td>
<td>0.048</td>
</tr>
<tr>
<td>PW⁴={t1,t5,t6,t3}</td>
<td>0.072</td>
</tr>
<tr>
<td>PW⁵={t2,t6,t4}</td>
<td>0.168</td>
</tr>
<tr>
<td>PW⁶={t2,t5,t6}</td>
<td>0.252</td>
</tr>
<tr>
<td>PW⁷={t6,t4,t3}</td>
<td>0.072</td>
</tr>
<tr>
<td>PW⁸={t5,t6,t3}</td>
<td>0.108</td>
</tr>
</tbody>
</table>

0.112 = 0.4 × 0.7 × 0.4 × 1.0

t2 and t3 never appear in the same possible world!
Possible Worlds and Rules

- Possible worlds are governed by rules

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Loc</th>
<th>Car Make</th>
<th>Plate No</th>
<th>Speed</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>11:45</td>
<td>L1</td>
<td>Honda</td>
<td>X-123</td>
<td>130</td>
</tr>
<tr>
<td>t2</td>
<td>11:50</td>
<td>L2</td>
<td>Toyota</td>
<td>Y-245</td>
<td>120</td>
</tr>
<tr>
<td>t3</td>
<td>11:35</td>
<td>L3</td>
<td>Toyota</td>
<td>Y-245</td>
<td>80</td>
</tr>
<tr>
<td>t4</td>
<td>12:10</td>
<td>L4</td>
<td>Mazda</td>
<td>W-541</td>
<td>90</td>
</tr>
<tr>
<td>t5</td>
<td>12:25</td>
<td>L5</td>
<td>Mazda</td>
<td>W-541</td>
<td>110</td>
</tr>
<tr>
<td>t6</td>
<td>12:15</td>
<td>L6</td>
<td>Nissan</td>
<td>L-105</td>
<td>105</td>
</tr>
</tbody>
</table>

\[ Rules : (t2 \oplus t3), (t4 \oplus t5) \{ (t1 \rightarrow t2) \} \]

A new rule

<table>
<thead>
<tr>
<th>World</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW^1={t1,t2,t6,t4}</td>
<td>0.16</td>
</tr>
<tr>
<td>PW^2={t1,t2,t5,t6}</td>
<td>0.24</td>
</tr>
<tr>
<td>PW^3={t2,t6,t4}</td>
<td>0.12</td>
</tr>
<tr>
<td>PW^4={t2,t5,t6}</td>
<td>0.18</td>
</tr>
<tr>
<td>PW^5={t6,t4,t3}</td>
<td>0.12</td>
</tr>
<tr>
<td>PW^6={t5,t6,t3}</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Correlation and Dependencies

• An example of correlated tuples

<table>
<thead>
<tr>
<th>TID</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.4</td>
</tr>
<tr>
<td>t2</td>
<td>0.42</td>
</tr>
<tr>
<td>t3</td>
<td>0.468</td>
</tr>
</tbody>
</table>

A probabilistic table

A factored representation:

\[
\Pr(t1 = x1, t2 = x2, t3 = x3) = f_1(t1 = x1)f_{12}(t1 = x1, t2 = x2)f_{23}(t2 = x2, t3 = x3)
\]
Possible Worlds

- Compute the joint probability of possible world assignments (Details in [Sen and Deshpande, ICDE’07])

<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>Pr(t1,t2,t3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.378</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.162</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.018</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.042</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.028</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.012</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.108</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.252</td>
</tr>
</tbody>
</table>

Joint probability of (t1,t2,t3)

<table>
<thead>
<tr>
<th>World</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW1=∅</td>
<td>0.378</td>
</tr>
<tr>
<td>PW2={t3}</td>
<td>0.162</td>
</tr>
<tr>
<td>PW3={t2}</td>
<td>0.018</td>
</tr>
<tr>
<td>PW4={t2,t3}</td>
<td>0.042</td>
</tr>
<tr>
<td>PW5={t1}</td>
<td>0.028</td>
</tr>
<tr>
<td>PW6={t1,t3}</td>
<td>0.012</td>
</tr>
<tr>
<td>PW7={t1,t2}</td>
<td>0.108</td>
</tr>
<tr>
<td>PW8={t1,t2,t3}</td>
<td>0.252</td>
</tr>
</tbody>
</table>

Possible worlds
Conceptual Query Answering

Adapted from Singh et al. ICDE’08
Attribute Level Uncertainty

• An aerial photograph of a battlefield

A friendly tank $a$

A friendly transport $b$

Unknown vehicle $d$

An enemy tank $c$

Antova et al. ICDE’08
Attribute Level Uncertainty

- A relation $R(\text{ID}, \text{Type}, \text{Faction})$ with uncertain attributes
  - $\text{ID} = \{1, 2, 3, 4\}$
  - $\text{Type} = \{\text{Tank, Transport}\}$
  - $\text{Faction} = \{\text{Friend, Enemy}\}$

- Uncertainty in data
  - Vehicle 1 is a friendly tank $a$
  - Vehicle 2 and 3 are either
    - a friendly transport $b$, or
    - an enemy tank $c$
  - Vehicle 4 is unknown vehicle $d$

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>b</td>
<td>?</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>c</td>
<td>?</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>d</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Representing Uncertainty

- **ID of vehicle b and c**
  - “b’s ID is 2 and c’s ID is 3”, or “b’s ID is 3 and c’s ID is 2”?
  - Random variable $x=\{1,2\}$

- **Type of Vehicle d**
  - “Tank” or “Transport”?
  - Random variable $y=\{1,2\}$

- **Faction of Vehicle d**
  - “Friend” or “Enemy”?
  - Random variable $z=\{1,2\}$

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>b</td>
<td>?</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>c</td>
<td>?</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>d</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
U-Relation

- Vertical Representation
  - Use a U-relation to represent each attribute of relation R

<table>
<thead>
<tr>
<th>D</th>
<th>Vehicle</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>x=1</td>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>3</td>
</tr>
<tr>
<td>x=2</td>
<td>b</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>4</td>
</tr>
</tbody>
</table>

U-relation for “ID”

<table>
<thead>
<tr>
<th>D</th>
<th>Vehicle</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>x=1</td>
<td>a</td>
<td>Tank</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Transport</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Tank</td>
</tr>
<tr>
<td>y=1</td>
<td>d</td>
<td>Tank</td>
</tr>
</tbody>
</table>

U-relation for “Type”

<table>
<thead>
<tr>
<th>D</th>
<th>Vehicle</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>Friend</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Friend</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Enemy</td>
</tr>
<tr>
<td>z=1</td>
<td>d</td>
<td>Friend</td>
</tr>
<tr>
<td>z=2</td>
<td>d</td>
<td>Enemy</td>
</tr>
</tbody>
</table>

U-relation for “Faction”
Possible Worlds of U-Relations

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>z</th>
<th>b.ID &amp; c.ID (x)</th>
<th>d.Type (y)</th>
<th>d.Faction(z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>b.ID=2, c.ID=3</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>b.ID=2, c.ID=3</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>b.ID=2, c.ID=3</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>b.ID=2, c.ID=3</td>
<td>Transport</td>
<td>Enemy</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>b.ID=3, c.ID=2</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>b.ID=3, c.ID=2</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>b.ID=3, c.ID=2</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>b.ID=3, c.ID=2</td>
<td>Transport</td>
<td>Enemy</td>
</tr>
</tbody>
</table>

Possible worlds
Transformation of U-Relation

- U-Relations can be transformed to a probabilistic table

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>b</td>
<td>?</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>c</td>
<td>?</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>d</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

b.ID=2, c.ID=3 (30%)
b.ID=3, c.ID=2 (70%)
d.Type=Tank(50%), Transport(50%)
d.Faction=Friend (50%), Enemy(50%)

<table>
<thead>
<tr>
<th>TID</th>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
<th>Conf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
<td>1</td>
</tr>
<tr>
<td>t2</td>
<td>b</td>
<td>2</td>
<td>Transport</td>
<td>Friend</td>
<td>0.3</td>
</tr>
<tr>
<td>t3</td>
<td>c</td>
<td>3</td>
<td>Tank</td>
<td>Enemy</td>
<td>0.3</td>
</tr>
<tr>
<td>t4</td>
<td>b</td>
<td>3</td>
<td>Tank</td>
<td>Enemy</td>
<td>0.7</td>
</tr>
<tr>
<td>t5</td>
<td>c</td>
<td>2</td>
<td>Transport</td>
<td>Friend</td>
<td>0.7</td>
</tr>
<tr>
<td>t6</td>
<td>d</td>
<td>4</td>
<td>Tank</td>
<td>Friend</td>
<td>0.25</td>
</tr>
<tr>
<td>t7</td>
<td>d</td>
<td>4</td>
<td>Tank</td>
<td>Enemy</td>
<td>0.25</td>
</tr>
<tr>
<td>t8</td>
<td>d</td>
<td>4</td>
<td>Transport</td>
<td>Friend</td>
<td>0.25</td>
</tr>
<tr>
<td>t9</td>
<td>d</td>
<td>4</td>
<td>Transport</td>
<td>Enemy</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Generation rules: t2→t3, t4→t5, t2⊕t4, t3⊕t5
t6⊕t7⊕t8⊕t9
Continuous Uncertain Model

- An attribute may take a continuous PDF as the value
- A table $T=(\Sigma_T, \Delta_T)$
  - $\Sigma_T$: a relational schema
  - $\Delta_T$: dependency information including pdfs or joint pdfs
  - For each dependent group of uncertain attributes, store history $\Lambda$. When a new tuple is added, check whether the dependency remains

<table>
<thead>
<tr>
<th>Car-id</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Gaussian(mean 18, variance 6)</td>
</tr>
<tr>
<td>C2</td>
<td>Uniform(center (32, 26), radius 7)</td>
</tr>
</tbody>
</table>
More on Possible Worlds

- The possible world model can be enriched by various kinds of (arbitrarily complicated) constraints
  - Example: if instances A.a and B.b appear, instances C.c or D.d must appear
- Completeness and closure
  - A model M is closed under an operation Op if applying Op on any uncertain relation in M results in an uncertain relation that can be represented in M
  - M is complete if M is closed for all relational operations
    - Completeness $\rightarrow$ closure, but not the other way
- More details in [Sarma et al. ICDE’06]
Summary

• Object/tuple level and attribute level uncertainty
• Possible worlds model
• Expressiveness
  – Should be closed under the application operations
  – Completeness is even better
• Succinctness: representing a large number of worlds using fairly little space
• Evaluation efficiency: complexity in useful queries
  – Often a tradeoff between succinctness and efficiency
• Ease of use: can be put on top of an RDBMS
• [Antova et al. ICDE’08]
Outline

• Uncertainty and uncertain data, where and why?
• Models for uncertain and probabilistic data
• (coffee break)
• OLAP on uncertain and probabilistic data
• Mining uncertain and probabilistic data
• Tools: querying uncertain and probabilistic data
  – Indexing uncertain and probabilistic data
  – Ranking queries and spatial queries
• Summary and discussion
OLAP Query

What are the total repair cost for F150’s in the East?

<table>
<thead>
<tr>
<th>Auto</th>
<th>Loc</th>
<th>Repair</th>
<th>Text</th>
<th>Brake</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>F-150</td>
<td>NY</td>
<td>$200</td>
<td>(0.8, 0.2)</td>
</tr>
<tr>
<td>p2</td>
<td>F-150</td>
<td>MA</td>
<td>$250</td>
<td>(0.9, 0.1)</td>
</tr>
<tr>
<td>p3</td>
<td>F-150</td>
<td>CA</td>
<td>$150</td>
<td>(0.7, 0.3)</td>
</tr>
<tr>
<td>p4</td>
<td>Sierra</td>
<td>TX</td>
<td>$300</td>
<td>(0.3, 0.7)</td>
</tr>
<tr>
<td>p5</td>
<td>Camry</td>
<td>TX</td>
<td>$325</td>
<td>(0.7, 0.3)</td>
</tr>
<tr>
<td>p6</td>
<td>Camry</td>
<td>TX</td>
<td>$175</td>
<td>(0.5, 0.5)</td>
</tr>
<tr>
<td>p7</td>
<td>Civic</td>
<td>TX</td>
<td>$225</td>
<td>(0.3, 0.7)</td>
</tr>
<tr>
<td>p8</td>
<td>Civic</td>
<td>TX</td>
<td>$120</td>
<td>(0.2, 0.8)</td>
</tr>
<tr>
<td>p9</td>
<td>F150</td>
<td>East</td>
<td>$140</td>
<td>(0.5, 0.5)</td>
</tr>
<tr>
<td>p10</td>
<td>Truck</td>
<td>TX</td>
<td>$500</td>
<td>(0.9, 0.1)</td>
</tr>
</tbody>
</table>
Three Options

• None: ignore all imprecise facts
  – Answer: p1, p2

• Contains: include only those contained in the query region
  – Answer: p1, p2, p9

• Overlaps: include all imprecise facts whose region overlaps the query region
  – Answer: p1, p2, p9, p10
Consistency among OLAP Queries

Q5 = Q3 + Q4 is expected!

Consistency does not hold for option contains, but holds for options none and overlaps!
Faithfulness of OLAP Queries

p9 is expected in Q5!

Faithfulness does not hold for option none, but holds for options contains and overlaps!
OLAP Requirements

- Consistency (summarizability): some natural relationships hold between answers to aggregation queries associated with different (connected) regions in a hierarchy
- Faithfulness: imprecise data should be considered properly in query answering
Possible Worlds
Allocation and Query Answering

• The allocation weights encode a set of possible worlds $D_1, \ldots, D_m$ with associated weights $w_1, \ldots, w_m$

• The answer to a query is a multiset $\{v_1, \ldots, v_m\}$

• Problem: how to summarize $\{v_1, \ldots, v_m\}$ properly?
Answer Variable

• Consider multiset \( \{v_1, \ldots, v_m\} \) of possible answers to a query \( Q \)

• Define the answer variable \( Z \) associated with \( Q \) to be a random variable with probability density function

\[
\Pr[Z=v_i] = \sum_{j \text{ s.t. } v_i=v_j} w_j, \quad 1 \leq i, j \leq m
\]
Answer Variable

- The answer to a query can be summarized as the first and the second moments (expected value and variance) of the answer variable Z.
- Basic faithfulness is satisfied if answers to queries are computed using the expected value of the answer variable.
Query Answering

- Identify the set of candidate facts and compute the corresponding allocations to Q
  - Identifying candidate facts: using a filter for the query region
  - Computing the corresponding allocations: identifying groups of facts that share the same identifier in the ID column, then summing up the allocations within each group
- Identify the information necessary to compute the summarization while circumventing the enumeration of possible worlds
Allocation Policies

- Dimension-independent allocation such as uniform allocation
- Measure-oblivious allocation such as count-based allocation
  - If Vancouver and Victoria have 100 and 50 F150’s, respectively, and there are another 30 in BC as imprecise records, then allocate 20 and 10 to Vancouver and Victoria, respectively
Outline

• Uncertainty and uncertain data, where and why?
• Models for uncertain and probabilistic data
• (coffee break)
• OLAP on uncertain and probabilistic data
• Mining uncertain and probabilistic data
• Tools: querying uncertain and probabilistic data
  – Indexing uncertain and probabilistic data
  – Ranking queries and spatial queries
• Summary and discussion
Probabilistic Transactions

• A transaction $t$ contains a number items where each item $x$ is associated with a positive probability $P_t(x)$
  – Assuming items in a transaction are independent
  – Itemset $xyz$ has probability $P_t(x)P_t(y)P_t(z)$ to happen in $t$

• In a probabilistic transaction database $D$ of $d$ transactions, an itemset $X$ is frequent if its expected support is at least $\rho d$, where $\rho$ is a user-specified support threshold
  – [Chui et al., PAKDD’07]
Possible Worlds of Transactions

- Enumerating all possible worlds to compute the expected supports is computationally infeasible for large transaction databases

Chui et al., PAKDD’07
Independent Transactions

• If transactions are independent, expected support can be calculated efficiently transaction by transaction

\[ S_e(X) = \sum_{j=1}^{d} \prod_{x \in X} P_{t_j}(x) \]

• Anti-monotonicity still holds: if X is infrequent, then every super set of X cannot be frequent

• U-Apriori: extending Apriori straightforwardly
Insignificant Support Contributions

• If a, b, c have existence probabilities 5%, 0.5%, and 0.1%, respectively in a transaction t, t contributes only 0.00000025 to the support of abc
  – In certain transactions, every transaction contributes 1 to the support of an itemset

• Counting many insignificant support contributions is costly
The Data Trimming Framework

- Obtain $D^T$ by removing the items with existential probabilities smaller than a trimming threshold $\rho_t$
  - $\rho_t$ can be either global to all items or local to each item
  - Estimate the error $e(X)$ in support counting introduced by reducing $D$ to $D^T$
- Mine $D^T$ using U-Apriori
  - If $X$ is frequent in $D^T$, $X$ must be frequent in $D$
  - If $X$ is infrequent in $D^T$, $X$ may or may not be infrequent in $D$
- If $\text{sup}_{D^T}(X) + e(X) < \rho d$, then $X$ can be pruned
  - Check supports for only those itemsets that cannot be pruned
Decremental Pruning

- Estimate upper bounds of candidate itemsets’ expected supports progressively when transactions are processed.
- If a candidate’s upper bound falls below the support threshold, the candidate can be pruned immediately.
- For $X' \subset X$, $k \geq 0$, $\text{sup}(X) \leq s(X, X', k)$, where

$$S(X, X', k) = \sum_{i=1}^{k} \prod_{x \in X} P_{t_i}(x) + \sum_{i=k+1}^{d} \prod_{x \in X'} P_{t_i}(x)$$

- Using singleton itemsets or prefix-sharing itemsets to compute $s(X, X', k)$ efficiently
  - Details in [Chui and Kao, PAKDD’08]
Is Expectation Good Enough?

- In D1, if the support threshold is 0.5, then a is frequent, however, a has only 50% chance to have support 0.5
- In D2, if the support threshold is 0.5, then a is infrequent. However, a has a probability of 0.9 to be frequent

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D1</strong></td>
<td><strong>Items</strong></td>
</tr>
<tr>
<td>TID</td>
<td>Items</td>
</tr>
<tr>
<td>t1</td>
<td>(a: 0.5)</td>
</tr>
<tr>
<td>t2</td>
<td>(b: 0.6)</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D2</strong></td>
<td><strong>Items</strong></td>
</tr>
<tr>
<td>TID</td>
<td>Items</td>
</tr>
<tr>
<td>t1</td>
<td>(a: 0.9), (b: 0.1)</td>
</tr>
<tr>
<td>t2</td>
<td>(c: 1)</td>
</tr>
</tbody>
</table>
Probabilistic Heavy Hitters

• An item is a \((\rho, \tau)\)-probabilistic heavy hitter if

\[
\sum_{w \in W, \sup_w(x) \geq \rho d} \Pr(w) \geq \tau
\]

– \(\tau\) is the probability/confidence threshold

• Dynamic programming using Poisson Binomial Recurrence

\[
B^t[0,0] = 1
\]

\[
B^t[i,0] = 0 \quad (i \geq 1)
\]

\[
B^t[i, j] = \begin{cases} 
B^t[i, j-1] & \text{if } w_j \neq t; \\
B^t[i, j-1](1 - p_j) + B^t[i-1, j-1]p_j & \text{if } w_j \neq t.
\end{cases}
\]

Zhang et al., SIGMOD’08
Classification on Uncertain Data

- Many studies exist in machine learning (particularly statistical learning)
  - Examples: [M. Mohri. Learning from Uncertain Data. COLT'03] and [S. Jain et al. Absolute Versus Probabilistic Classification in a Logical Setting. ALT'05]

- New problem: how does uncertain data affect classification?
  - How can we apply the existing classification with minor revision on uncertain data?
1NN Classification on Certain Data

- Point $x$ will be classified using point $y$ since $\text{dist}(x, y) < \text{dist}(x, z)$
1NN on Uncertain Data

- Object x may have a good chance to be classified using z
  - Instances of x have a high probability to lie in the error boundary of z
- When classification on uncertain data, it is important to use the relative errors of different data points over the different dimensions in order to improve the accuracy

Aggarwal ICDE’07
Density Estimation with Errors

- Kernel estimation
  - General form
    \[ \hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K_h'(x - \overline{X}_i) \]

  - Gaussian kernel with width \( h \)
    \[ \hat{f}(x) = \frac{1}{Nh\sqrt{2\pi}} \sum_{i=1}^{N} e^{-\frac{(x-\overline{X}_i)^2}{2h^2}} \]

- Error at point \( \overline{X}_i \) can be modeled by function \( \psi(\overline{X}_i) \)

- Error-based kernel
  \[
  Q_h'(x - \overline{X}_i, \psi(\overline{X}_i)) = \frac{h + \psi(\overline{X}_i)}{\sqrt{2\pi}} e^{-\frac{(x-\overline{X}_i)^2}{2(h^2 + \psi(\overline{X}_i)^2)}}
  \]

  \[
  f^Q(x, \psi(\overline{X}_i)) = \frac{1}{N} \sum_{i=1}^{N} Q_h'(x - \overline{X}_i, \psi(\overline{X}_i))
  \]

J. Pei, M. Hua, Y. Tao, and X. Lin: Mining Uncertain and Probabilistic Data
Error-Based Micro-Clustering

• Applying density estimation with errors on a large database may be costly

• Use micro-clusters to approximate
  – A BIRCH-like method [Zhang et al., SIGMOD’96]
  – Use the framework in [Aggarwal et al., VLDB’03], but maintain only q randomly chosen centroids
  – When assigning a point into a micro-cluster, use error-adjusted distance

\[
    dist(\overline{X}, c) = \sum_{i=1}^{d} \max \{0, (X_i - c_i)^2 - \psi_i(\overline{X})^2\}
\]

• Micro-clusters can be used to generate classification rules
Fuzzy Clustering

• Each data point is certain
• Clusters are fuzzy (uncertain to some extent)
  – No sharp boundary between clusters, often perform better in some applications
  – Each point is assigned to a cluster with a probability (membership degree)
• Hoppner et al. Fuzzy cluster analysis. Wiley, 1999
Clustering Multi-represented Objects

• An object may have multiple representations
  – Molecules are characterized by an amino acid sequence, a secondary structure and a 3D representation

• Clustering multi-represented objects needs to consider all representations in question
  – Combine distance/neighborhoods in all representations into one global distance/neighborhood
Clustering Uncertain Objects

• Objects are fuzzy/uncertain, clusters can be certain or fuzzy
  – A fuzzy object can be represented by a probability density function or a set of instances
  – All instances of an object are in the same space, different objects may have a different number of instances

• In clustering, the distribution of the distance between two objects and the probability that an object is a cluster center should be considered

\[
\Pr[a \leq \text{dist}(o, o') \leq b] = \int_{a}^{b} \Pr[\text{dist}(o, o') = x]dx
\]

Kriegel and Pfeifle, KDD’05, ICDM’05
K-means on Uncertain Data

- Run k-means, use expectation of distance to assign objects/probabilistic points to clusters
- Computation can be sped up by using bounding rectangles or other polygon to bound PDF regions and approximate distance calculation
Example

- $O_i$ cannot be assigned to $p_3$

Ngai et al., ICDM’06
(\(\alpha\), \(\beta\))-bicriteria Approximation

- Optimal k-center, k-means, and k-median are NP-hard even for certain data
- A (\(\alpha\), \(\beta\))-bicriteria approximation to k-clustering finds a clustering of size \(\beta k\) whose cost is at most \(\alpha\) times the cost of the optimal k-clustering
- Assigned clustering: an object is assigned to a cluster
- Unassigned clustering: only cluster centers are computed – different instances of an object may be assigned to different clusters
Theoretical Results

Cormode and McGregor, PODS’08

<table>
<thead>
<tr>
<th>Objective</th>
<th>Metric</th>
<th>Assignment</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-center Point probability</td>
<td>Any</td>
<td>Unassigned</td>
<td>$1 + \varepsilon$</td>
<td>$O(\varepsilon^{-1} \log^2 n)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$12 + \varepsilon$</td>
<td>$2$</td>
</tr>
<tr>
<td>K-center Discrete PDF</td>
<td>Any</td>
<td>Unassigned</td>
<td>$1.582 + \varepsilon$</td>
<td>$O(\varepsilon^{-1} \log^2 n)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$18.99 + \varepsilon$</td>
<td>$2$</td>
</tr>
<tr>
<td>K-means</td>
<td>Euclidean</td>
<td>Unassigned</td>
<td>$1 + \varepsilon$</td>
<td>$1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$3 + \varepsilon$</td>
<td></td>
</tr>
<tr>
<td>K-median</td>
<td>Any</td>
<td>Unassigned</td>
<td>$3 + \varepsilon$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Euclidean</td>
<td></td>
<td>$1 + \varepsilon$</td>
<td>$1$</td>
</tr>
<tr>
<td></td>
<td>Assigned</td>
<td></td>
<td>$7 + \varepsilon$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Euclidean</td>
<td>Assigned</td>
<td>$3 + \varepsilon$</td>
<td></td>
</tr>
</tbody>
</table>
K-center \((1 + \varepsilon)\) Approximation

- For each point \(x\), assign a weight
  \[ w_x = -\ln (1 - p_x) \]
- Greedily select a set of centers
  - Suppose \(c_1, \ldots, c_i\) are the current centers
  - A point \(x\) is assigned to a current cluster if it is within distance \(r\) to the center
  - Among the remaining points, find a new center \(c_{i+1}\) such that the total weight of points that can be assigned to \(c_{i+1}\) is maximized
Fuzzy Clustering of Uncertain Data

• Data points are probabilistic
• Clusters are fuzzy – each probabilistic point has a membership degree (between 0 and 1) to be assigned to a cluster
• Expectation maximization (EM) based on clustering of uncertain data [Dempster et al., J. of the Royal Stat. Society, 1977]
Outliers in Uncertain Data

- Which one is more an outlier, x or y?
Outlier Detection on Uncertain Data

• The $\eta$-probability of a data point is the probability that it lies in a region with data density at least $\eta$

• $(\delta, \eta)$-outlier: the $\eta$-probability of a point is some subspace is less than $\delta$

• Enumerate all non-empty subspaces in a bottom-up breadth-first search, for each subspace, check whether there is any $(\delta, \eta)$-outlier
  – Use sampling and micro-clusters to estimate density distribution
  – Details in [Aggarwal and Yu, SDM’08]
Outline

• Uncertainty and uncertain data, where and why?
• Models for uncertain and probabilistic data
• (coffee break)
• OLAP on uncertain and probabilistic data
• Mining uncertain and probabilistic data
• Tools: querying uncertain and probabilistic data
  – Indexing uncertain and probabilistic data
  – Ranking queries and spatial queries
• Summary and discussion
Conceptual Query Answering

Adapted from Singh et al. ICDE’08
U-Tree: Motivation

- Probabilistic range queries
  - Given query region $q$ and probability threshold $\tau$, return all the objects whose probability of being in $q$ is higher than $\tau$

- Appearance probability
  \[
  \Pr(C \text{ is in } q) = \int_{q \cap C} f(x) dx = \frac{\text{Area}(q \cap C)}{\text{Area}(C)}
  \]
  where $f(x)$ is the pdf of $C$
U-Tree: Idea

- Partition the object into three parts in one dimension (horizontally)
- Partition the object into three parts in the other dimension (vertically)
Range query \( q \), Prob. threshold \( \tau = 0.8 \)

- \( \Pr(\text{O is in } q) < \tau \), because \( q \) is disjoint with the right part of \( L_{1+} \), whose probability is \( p = 0.2 \)
- Thus, O can be pruned
U-Tree: Validation

- Pr(O is in q) > \tau, since q fully covers the part of O on the right side of L_{1+}, whose probability is p=0.2
- Thus, O can be validated
U-Tree: What to Store?

- Probabilistic constraint region (PCR)
- Select $0 < p_1 < \ldots < p_m < 0.5$, and compute $\text{PCR}(p_1), \ldots, \text{PCR}(p_m)$

![Diagram showing uncertain object O and information stored in a U-tree.](image)
U-Tree

- Structure
  - Root
  - Intermediate entry: A set of objects $O_1, \ldots, O_k$
    - Pointer to its child nodes
    - MBR of the PCR’s of $O_1, \ldots, O_k$
  - Leaf entry: Object $O$
    - Disk address of $O$
    - MBR of $O$
    - A set of PCR’s of $O$

- Query evaluation
Probabilistic Categorical Data

- Uncertain attribute “Problem”: derived from a text classifier
- Probabilistic threshold queries:
  - Find the tuples whose problem is “Brake” with probability 0.3 and “Tires” with probability 0.7
  - \( q = \{(\text{Brake}, 0.3), (\text{Tire}, 0.7)\} \)
  - \( \Pr(t_1.\text{Problem}=q) = 0.3 \times 0.5 + 0.5 \times 0.7 = 0.5 \)

<table>
<thead>
<tr>
<th>Make</th>
<th>Location</th>
<th>Date</th>
<th>Text</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>Explorer</td>
<td>WA</td>
<td>2/3/06</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>Camry</td>
<td>CA</td>
<td>3/5/05</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( t_3 )</td>
<td>Civic</td>
<td>TX</td>
<td>10/2/06</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( t_4 )</td>
<td>Caravan</td>
<td>IN</td>
<td>7/2/06</td>
<td>( \cdots )</td>
</tr>
</tbody>
</table>

\( \{(\text{Brake}, 0.5), (\text{Tires}, 0.5)\} \)
\( \{(\text{Trans}, 0.2), (\text{Suspension}, 0.8)\} \)
\( \{(\text{Exhaust}, 0.4), (\text{Brake}, 0.6)\} \)
\( \{(\text{Trans}, 1.0)\} \)
Probabilistic Inverted Index

<table>
<thead>
<tr>
<th>Make</th>
<th>Location</th>
<th>Date</th>
<th>Text</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explorer</td>
<td>WA</td>
<td>2/3/06</td>
<td>⋮</td>
<td>{(Brake, 0.5), (Tires, 0.5)}</td>
</tr>
<tr>
<td>Camry</td>
<td>CA</td>
<td>3/5/05</td>
<td>⋮</td>
<td>{(Trans, 0.2), (Suspension, 0.8)}</td>
</tr>
<tr>
<td>Civic</td>
<td>TX</td>
<td>10/2/06</td>
<td>⋮</td>
<td>{(Exhaust, 0.4), (Brake, 0.6)}</td>
</tr>
<tr>
<td>Caravan</td>
<td>IN</td>
<td>7/2/06</td>
<td>⋮</td>
<td>{(Trans, 1.0)}</td>
</tr>
</tbody>
</table>

A list of domain element

- Brake: (t₃, 0.6) (t₁, 0.5)
- Tire: (t₁, 0.5)
- Trans: (t₅, 1) (t₂, 0.2)
- Suspension: (t₂, 0.8)
- Exhaust: (t₄, 0.4)

In probability descending order
Query Answering

• On Attribute A, a query \( q=\{(d_3,0.4),(d_6,0.1),(d_8,0.2)\} \), \( \tau=0.3 \)
  – \( \Pr(q=t.A)= p'_3 \times 0.4 + p'_6 \times 0.1 + p'_8 \times 0.2 \)

• Column pruning
  – If for each \( d_i \in t.A \), \( \Pr(d_i) < \tau \),
    then \( t \) can be pruned

• Row pruning
  – If \( t \) only contains \( d_6 \) and \( d_8 \) whose probability is smaller than \( \tau \),
    then the \( t \) can be pruned
Ranking Queries

• Find the top-2 sensors with highest temperature
  – Certain data: answer = \{R1, R2\}
  – Uncertain data
    • R1 and R2 may not co-exist in a possible world
    • In different possible worlds, the answers are different

<table>
<thead>
<tr>
<th>RID</th>
<th>Loc.</th>
<th>Time</th>
<th>Sensor-id</th>
<th>Temperature</th>
<th>Conf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>A</td>
<td>6/2/06 2:14</td>
<td>S101</td>
<td>25</td>
<td>0.3</td>
</tr>
<tr>
<td>R2</td>
<td>B</td>
<td>7/3/06 4:07</td>
<td>S206</td>
<td>21</td>
<td>0.4</td>
</tr>
<tr>
<td>R3</td>
<td>B</td>
<td>7/3/06 4:09</td>
<td>S231</td>
<td>13</td>
<td>0.5</td>
</tr>
<tr>
<td>R4</td>
<td>A</td>
<td>4/12/06 20:32</td>
<td>S101</td>
<td>12</td>
<td>1.0</td>
</tr>
<tr>
<td>R5</td>
<td>E</td>
<td>3/13/06 22:31</td>
<td>S063</td>
<td>17</td>
<td>0.8</td>
</tr>
<tr>
<td>R6</td>
<td>E</td>
<td>3/13/06 22:28</td>
<td>S732</td>
<td>11</td>
<td>0.2</td>
</tr>
</tbody>
</table>

R2 ⊕ R3  R5 ⊕ R6
Challenges

• What does a probabilistic ranking query mean?
  – A ranking query on certain data returns the best k results in the ranking function
  – Ranking queries on uncertain data may be formulated differently to address different application interests

• How can a ranking query be answered efficiently?
  – Answering ranking queries on probabilistic databases can be very costly when the number of possible worlds is huge
Query Types

• How are tuples ranked?

Ranking based on objective functions and output probabilities: Global-Topk
Ranking Based on Objective Functions

- A scoring function is given
  - Rank the sensors in temperature descending order and select the top-2 results
    \[ R1 < R2 < R5 < R3 < R4 < R1 \]
- How should the top-2 ranking results be captured?

<table>
<thead>
<tr>
<th>RID</th>
<th>Loc.</th>
<th>Time</th>
<th>Sensor-id</th>
<th>Temperature</th>
<th>Conf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>A</td>
<td>6/2/06 2:14</td>
<td>S101</td>
<td>25</td>
<td>0.3</td>
</tr>
<tr>
<td>R2</td>
<td>B</td>
<td>7/3/06 4:07</td>
<td>S206</td>
<td>21</td>
<td>0.4</td>
</tr>
<tr>
<td>R3</td>
<td>B</td>
<td>7/3/06 4:09</td>
<td>S231</td>
<td>13</td>
<td>0.5</td>
</tr>
<tr>
<td>R4</td>
<td>A</td>
<td>4/12/06 20:32</td>
<td>S101</td>
<td>12</td>
<td>1.0</td>
</tr>
<tr>
<td>R5</td>
<td>E</td>
<td>3/13/06 22:31</td>
<td>S063</td>
<td>17</td>
<td>0.8</td>
</tr>
<tr>
<td>R6</td>
<td>E</td>
<td>3/13/06 22:28</td>
<td>S732</td>
<td>11</td>
<td>0.2</td>
</tr>
</tbody>
</table>

\[ R2 \oplus R3 \quad R5 \oplus R6 \]
U-Topk Queries

- Find the most probable top-2 list in possible worlds
  - \( \langle R1, R2 \rangle \): p=0.12
  - \( \langle R1, R5 \rangle \): p=0.144
  - \( \langle R1, R3 \rangle \): p=0.03
  - \( \langle R1, R4 \rangle \): p=0.006
  - \( \langle R2, R5 \rangle \): p=0.224
  - \( \langle R2, R4 \rangle \): p=0.056
  - \( \langle R5, R3 \rangle \): p=0.28
  - \( \langle R3, R4 \rangle \): p=0.07
  - \( \langle R5, R4 \rangle \): p=0.056
  - \( \langle R4, R6 \rangle \): p=0.014

- Answer: \( \langle R5, R3 \rangle \)
U-kRanks Queries

- Find the tuple of the highest probability at each ranking position
  - **The 1st position**
    - R1: \( p=0.3 \)
    - R2: \( p=0.28 \)
    - R5: \( p=0.336 \)
    - R3: \( p=0.07 \)
    - R4: \( p=0.014 \)
  - **The 2nd position**
    - R5: \( p=0.368 \)

- Answer: \( \langle R5, R5 \rangle \)
PT-k Queries

• Find the tuples whose probabilities to be in the top-2 list are at least \( p \) (\( p = 0.35 \))
  – R1: \( p = 0.3 \)
  – R2: \( p = 0.4 \)
  – R3: \( p = 0.38 \)
  – R4: \( p = 0.202 \)
  – R5: \( p = 0.704 \)
  – R6: \( p = 0.014 \)

• Answer: \{R2,R3,R5\}

<table>
<thead>
<tr>
<th>Possible world</th>
<th>Probability</th>
<th>Top-2 on Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>( {R_1, R_2, R_4, R_5} )</td>
<td>0.096</td>
<td>( R_1, R_2 )</td>
</tr>
<tr>
<td>( {R_1, R_2, R_4, R_6} )</td>
<td>0.024</td>
<td>( R_1, R_2 )</td>
</tr>
<tr>
<td>( {R_1, R_3, R_4, R_5} )</td>
<td>0.12</td>
<td>( R_1, R_5 )</td>
</tr>
<tr>
<td>( {R_1, R_3, R_4, R_6} )</td>
<td>0.03</td>
<td>( R_1, R_3 )</td>
</tr>
<tr>
<td>( {R_1, R_4, R_5} )</td>
<td>0.024</td>
<td>( R_1, R_5 )</td>
</tr>
<tr>
<td>( {R_1, R_4, R_6} )</td>
<td>0.006</td>
<td>( R_1, R_4 )</td>
</tr>
<tr>
<td>( {R_2, R_4, R_5} )</td>
<td>0.224</td>
<td>( R_2, R_5 )</td>
</tr>
<tr>
<td>( {R_2, R_4, R_6} )</td>
<td>0.056</td>
<td>( R_2, R_4 )</td>
</tr>
<tr>
<td>( {R_3, R_4, R_5} )</td>
<td>0.28</td>
<td>( R_5, R_3 )</td>
</tr>
<tr>
<td>( {R_3, R_4, R_6} )</td>
<td>0.07</td>
<td>( R_3, R_4 )</td>
</tr>
<tr>
<td>( {R_4, R_5} )</td>
<td>0.056</td>
<td>( R_5, R_4 )</td>
</tr>
<tr>
<td>( {R_4, R_6} )</td>
<td>0.014</td>
<td>( R_4, R_6 )</td>
</tr>
</tbody>
</table>
Global-Topk

- Find the top-2 tuples whose probabilities to be in the top-2 list are the highest
- Ranking based on objective functions and output probabilities

Example
- R1: p=0.3
- R2: p=0.4
- R3: p=0.38
- R4: p=0.202
- R5: p=0.704
- R6: p=0.014

Answer={R5,R2}

<table>
<thead>
<tr>
<th>Possible world</th>
<th>Probability</th>
<th>Top-2 on Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1 = {R1, R2, R4, R5}</td>
<td>0.096</td>
<td>R1, R2</td>
</tr>
<tr>
<td>W2 = {R1, R2, R4, R6}</td>
<td>0.024</td>
<td>R1, R2</td>
</tr>
<tr>
<td>W3 = {R1, R3, R4, R5}</td>
<td>0.12</td>
<td>R1, R5</td>
</tr>
<tr>
<td>W4 = {R1, R3, R4, R6}</td>
<td>0.03</td>
<td>R1, R3</td>
</tr>
<tr>
<td>W5 = {R1, R4, R5}</td>
<td>0.024</td>
<td>R1, R5</td>
</tr>
<tr>
<td>W6 = {R1, R4, R6}</td>
<td>0.006</td>
<td>R1, R4</td>
</tr>
<tr>
<td>W7 = {R2, R4, R5}</td>
<td>0.224</td>
<td>R2, R5</td>
</tr>
<tr>
<td>W8 = {R2, R4, R6}</td>
<td>0.056</td>
<td>R2, R4</td>
</tr>
<tr>
<td>W9 = {R3, R4, R5}</td>
<td>0.28</td>
<td>R5, R3</td>
</tr>
<tr>
<td>W10 = {R3, R4, R6}</td>
<td>0.07</td>
<td>R3, R4</td>
</tr>
<tr>
<td>W11 = {R4, R5}</td>
<td>0.056</td>
<td>R5, R4</td>
</tr>
<tr>
<td>W12 = {R4, R6}</td>
<td>0.014</td>
<td>R4, R6</td>
</tr>
</tbody>
</table>
Query Answering Methods

- The dominant set property
  - For any tuple $t$, whether $t$ is in the answer set only depends on the tuples ranked higher than $t$
  - The dominant set of $t$ is the subset of tuples in $T$ that are ranked higher than $t$
    - E.g. the dominant set of $R3$ is $S_{R3} = \{R1, R2, R5\}$

- Framework of Query Answering Methods
  - Retrieve tuples in the ranking order
  - Evaluate each tuple based on its dominant set

<table>
<thead>
<tr>
<th>Ranked tuples:</th>
<th>Temperature</th>
<th>25</th>
<th>21</th>
<th>17</th>
<th>13</th>
<th>12</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>RID</td>
<td>R1</td>
<td>R2</td>
<td>R5</td>
<td>R3</td>
<td>R4</td>
<td>R6</td>
<td></td>
</tr>
</tbody>
</table>
Answering PT-k Queries

- Position probability \( \Pr(t_i,j) \)
  - The probability that \( t_i \) is ranked at the \( j \)-th position
  - E.g. \( \Pr(R3,2) = \Pr(R3) \times \Pr(S_{R3},1) \)

<table>
<thead>
<tr>
<th>Temperature</th>
<th>25</th>
<th>21</th>
<th>17</th>
<th>13</th>
<th>12</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>RID</td>
<td>R1</td>
<td>R2</td>
<td>R5</td>
<td>R3</td>
<td>R4</td>
<td>R6</td>
</tr>
</tbody>
</table>

R3 is ranked 2\(^{nd}\), if R3 appears, and 1 tuple in \( S_{R3} \) appears

- Generally: \( \Pr(t_i, j) = \Pr(t_i) \times \Pr(S_{t_i}, j-1) \)
Answering PT-k Queries

- Subset probability $\Pr(S_{ti}, j)$
  - The probability that $j$ tuples appear in $S_{ti}$
  - E.g. $S_{R3} = \{R5\} \cup S_{R5}$
  - $\Pr(S_{R3}, 2) = \Pr(R5) \times \Pr(S_{R5}, 1) + (1 - \Pr(R5)) \times \Pr(S_{R5}, 2)$

  $\begin{array}{|c|c|c|c|c|c|c|c|}
  \hline
  \text{Temperature} & 25 & 21 & 17 & 13 & 12 & 11 \\
  \text{RID} & R1 & R2 & R5 & R3 & R4 & R6 \\
  \hline
  \end{array}$

  2 tuples appear in $S_{R3}$, if \{ $R5$ appears, 1 tuple appears in $S_{R5}$ \}
  \{ $R5$ does not appear, 2 tuples appear in $S_{R5}$ \}

- Generally (Poisson Binomial Recurrence):
  \[ \Pr(S_{ti}, j) = \Pr(t_i) \times \Pr(S_{ti-1}, j-1) + (1 - \Pr(t_i)) \times \Pr(S_{ti-1}, j) \]
Summary of Query Answering Methods

- Optimal algorithms for U-Topk and U-kRanks queries in terms of the number of accessed tuples (Soliman et al. ICDE’07)
- Query answering algorithms for U-Topk and U-kRanks queries based on Poisson binomialial recurrence (Yi et al. ICDE’08)
- Spatial and probabilistic pruning techniques for U-kRanks queries (Lian and Chen, EDBT’08)
- Efficient query answering algorithms and pruning techniques for PT-k queries (Hua et al. ICDE’08, SIGMOD’08)
- A sampling-based method (Silberstein et al. ICDE’06)
Ranking Based on Output Probabilities

- **Query Q**: find the average temperature of all sensors
- **Ranking**: find the top-2 results with the highest probabilities of being the answers to Q (output probabilities)
  - Answer: 14 (p=0.28), 16.67 (p=0.224)

<table>
<thead>
<tr>
<th>Possible world</th>
<th>Probability</th>
<th>Average temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_1 = {R_1, R_2, R_4, R_5}$</td>
<td>0.096</td>
<td>18.75</td>
</tr>
<tr>
<td>$W_2 = {R_1, R_2, R_4, R_6}$</td>
<td>0.024</td>
<td>17.25</td>
</tr>
<tr>
<td>$W_3 = {R_1, R_3, R_4, R_5}$</td>
<td>0.12</td>
<td>16.75</td>
</tr>
<tr>
<td>$W_4 = {R_1, R_3, R_4, R_6}$</td>
<td>0.03</td>
<td>15.25</td>
</tr>
<tr>
<td>$W_5 = {R_1, R_4, R_5}$</td>
<td>0.024</td>
<td>18.00</td>
</tr>
<tr>
<td>$W_6 = {R_1, R_4, R_6}$</td>
<td>0.006</td>
<td>16.00</td>
</tr>
<tr>
<td>$W_7 = {R_2, R_4, R_5}$</td>
<td>0.224</td>
<td>16.67</td>
</tr>
<tr>
<td>$W_8 = {R_2, R_4, R_6}$</td>
<td>0.056</td>
<td>14.67</td>
</tr>
<tr>
<td>$W_9 = {R_3, R_4, R_5}$</td>
<td>0.28</td>
<td>14.00</td>
</tr>
<tr>
<td>$W_{10} = {R_3, R_4, R_6}$</td>
<td>0.07</td>
<td>12.00</td>
</tr>
<tr>
<td>$W_{11} = {R_4, R_5}$</td>
<td>0.056</td>
<td>14.50</td>
</tr>
<tr>
<td>$W_{12} = {R_4, R_6}$</td>
<td>0.014</td>
<td>11.50</td>
</tr>
</tbody>
</table>
Query Answering

• Monte Carlo Simulation (1 step)
  – Choose a possible world at random, and evaluate the query
  – Record the answer to the query and its frequency
• For example, if we run 100 steps of Monte Carlo simulation, and “14” is the answer in 30 steps
  – The output probability of “14” can be approximated by 30/100=0.3, with an error bound $\varepsilon$
  – The output probability of “14” lies in the probability interval $[0.3-\varepsilon, 0.3+\varepsilon]$
  – The more steps of Monte Carlo simulation we run, the smaller probability intervals we can get
Query Answering (cont.)

- The simulation stops when the top-k output probabilities and their relative ranks are clear
  - E.g. There are 5 possible results G1, G2, G3, G4 and G5. After a few steps of Monte Carlo simulation, the output probability interval of each result is shown below
  - G3’s output probability is in top-2. The other answer might be one of G1, G2, and G4. But G5’s output probability cannot be in top-2

Re et al. ICDE’07
More on Monte Carlo Simulation

- Separate data schema and uncertain variables
  - Data schema is certain
  - Use random variables supported by variable generation (VG) functions to simulate uncertainty
- A naïve implementation: run Monte Carlo simulation until the result is stable
- Efficient Implementation
  - Run N times of Monte Carlo simulation once in batch
  - Delay random attribute materialization as long as possible
  - Reproduce values for random attributes when necessary
- Details in [Jampani et al., SIGMOD’08]
Probabilistic Skyline

- Probabilistic skylines
  - An instance has a probability to represent the object
  - An object has a probability to be in the skyline
Skyline Probabilities

- Skyline probability
  - B is in the skyline of possible worlds \( w_1 = \{a_1, b_1, c_1\}, w_2 = \{a_1, b_1, c_2\}, w_3 = \{a_1, b_2, c_1\}, \) and \( w_4 = \{a_1, b_2, c_2\} \)
  - Thus, \( \Pr(B) = \Pr(w_1) + \Pr(w_2) + \Pr(w_3) + \Pr(w_4) = 4 \times 0.125 = 0.5 \)
- \( p \)-skyline = \( \{ U \mid \Pr(U) \geq p \} \) for a given threshold \( p \)
Probabilistic Skyline Computation

- **Iteration**: Bounding-Pruning-Refining
- **Bounding**
  - Bound \( Pr(u) \): lower bound \( Pr^-(u) \) and upper bound \( Pr^+(u) \)
  - Bound \( Pr(U) \): \( Pr(U) = \frac{1}{|U|} \sum_{u \in U} Pr(u) \)
- **Pruning**
  - In \( p \)-skyline if lower bound \( Pr^-(U) \geq p \)
  - Not in \( p \)-skyline if upper bound \( Pr^+(U) < p \)
- **Refining**
  - Bottom-up method
  - Top-down method
Bottom-up Method

- Key Idea
  - Two instances \( u_1 \) and \( u_2 \in U \), if \( u_1 \) dominants \( u_2 \), then \( \Pr(u_1) \geq \Pr(u_2) \)

- The layered structure
  - Sort the instances of an object according to the dominance relation

- Bounding
  - \( \max\{\Pr(u_1), \Pr(u_2)\} \geq \max\{\Pr(u_3), \Pr(u_4)\} \geq \Pr(u_5) \)
Top-down Method

- **Bounding**
  - Using the lower corner and upper corner to bound the skyline probability
  - \( \Pr(N_{min}) \leq \Pr(u) \leq \Pr(N_{max}) \)

- **Iterative partitioning**: binary tree
Outline

• Uncertainty and uncertain data, where and why?
• Models for uncertain and probabilistic data
• (coffee break)
• OLAP on uncertain and probabilistic data
• Mining uncertain and probabilistic data
• Tools: querying uncertain and probabilistic data
  – Indexing uncertain and probabilistic data
  – Ranking queries and spatial queries
• Summary and discussion
Summary

• Uncertain data becomes more and more important and prevalent
  – Critical applications: sensor networks, location-based services, web applications, user preferences, health-informatics, …

• Modeling uncertain data
  – Model uncertainty at various levels
  – Model correlation among data entries

• OLAP on uncertain data

• Mining uncertain data

• Tools: querying uncertain data
  – Simple queries, ranking queries, spatial queries
  – Using indexes to speed up query answering
Can Uncertainty Be Beneficial?

• In all the cases discussed so far, uncertainty leads to more complicated processing 😞

• Uncertainty and privacy preservation
  – Privacy preservation – preventing individuals from being re-identified, while keeping the aggregate data useful
  – Major approaches: perturbation and generalization – making data uncertain!

• [Aggarwal, ICDE’08]
Thank You

Future is uncertain because it will be what we make it.

– Immanuel Wallerstein
References (1)

• Parag Agrawal and Jennifer Widom. Confidence-aware joins in large uncertain databases. Technical report, Stanford University CA, USA.
References(2)

References (3)

References (4)


• Chun-Kit Chui and Ben Kao. A decremental approach for mining frequent itemsets from uncertain data. In PAKDD, pages 64-75, 2008.


References (5)

References(6)

• Xiang Lian and Lei Chen. Monochromatic and bichromatic reverse skyline search over uncertain databases. In SIGMOD, Vancouver, Canada, 2008.
References (7)

References(8)

- Anish Das Sarma, Martin Theobald, and Jennifer Widom. Exploiting lineage for confidence computation in uncertain and probabilistic databases. Technical report, Stanford University CA, USA.
References (9)

References (10)

