Foundations of Uncertain-Data Integration

Parag Agrawal

Joint work with Anish Das Sarma, Jeffrey D. Ullman, Jennifer Widom
Why integrate uncertain data?
Why integrate uncertain data?

- Uncertain databases
  - Store and query uncertain data
  - Sensor data, scientific data, extraction
Why integrate uncertain data?

- Uncertain databases
  - Store and query uncertain data
  - Sensor data, scientific data, extraction
- Data integration
  - Combines certain data from multiple sources
Why integrate uncertain data?

- Uncertain databases
  - Store and query uncertain data
  - Sensor data, scientific data, extraction
- Data integration
  - Combines certain data from multiple sources
- Uncertain data integration
  - Multiple uncertain data sources
Why integrate uncertain data?

- Uncertain databases
  - Store and query uncertain data
  - Sensor data, scientific data, extraction
- Data integration
  - Combines certain data from multiple sources
- Uncertain data integration
  - Multiple uncertain data sources
  - Uncertainty introduced during integration
Outcome of data integration
Outcome of data integration

- Certain data
Outcome of data integration

- Certain data
- More data
Outcome of data integration

- Certain data
  - More data
- Uncertain data

Possible Worlds
Outcome of data integration

- Certain data
  - More data

- Uncertain data
  - More data

Possible Worlds
Outcome of data integration

- Certain data
  - More data

- Uncertain data
  - More data
  - Less uncertainty
Examples

- Object, Location

2, A  2, B
Examples

- Object, Location

\[2, A \quad 2, B \quad 2, B \quad 2, C\]
Examples

- Object, Location

- 2, A
- 2, B
- 2, B
- 2, C
Examples

- Object, Location

<table>
<thead>
<tr>
<th>1, X</th>
<th>1, Y</th>
<th>2, B</th>
</tr>
</thead>
<tbody>
<tr>
<td>2, A</td>
<td>2, B</td>
<td>2, B</td>
</tr>
<tr>
<td>2, A</td>
<td>2, B</td>
<td>2, B</td>
</tr>
<tr>
<td>2, B</td>
<td>2, B</td>
<td>2, C</td>
</tr>
<tr>
<td>3, P</td>
<td>3, Q</td>
<td>2, C</td>
</tr>
</tbody>
</table>
Examples

- **Object, Location**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1, X</td>
<td>1, Y</td>
<td>2, B</td>
<td>2, C</td>
</tr>
<tr>
<td>2, A</td>
<td>2, B</td>
<td>3, P</td>
<td>3, Q</td>
</tr>
<tr>
<td>2, B</td>
<td>2, B</td>
<td>1, Y</td>
<td>2, B</td>
</tr>
<tr>
<td>2, C</td>
<td>2, C</td>
<td>2, A</td>
<td>3, P</td>
</tr>
</tbody>
</table>

Note: Some combinations are marked with an 'X'.
Examples

- Object, Location

Corroboration
Examples

- Object, Location

Contradiction

- 1, X
- 2, A
- 1, Y
- 2, B

Corroboration

- 2, A
- 2, B
- 2, B
- 2, B
- 2, C
- 2, C

1, Y
2, B
3, P
1, Y
2, B
3, P
Examples

- Object, Location

Contradiction

Exploit redundancy

Track correlations
Local-as-view data integration

$S_1$, $S_2$, ..., $S_k$
Local-as-view data integration

- Data sources

\[ S_1 \quad S_2 \quad \ldots \quad S_k \]
Local-as-view data integration

- Data sources
- Mediated schema
Local-as-view data integration

- Data sources
- Mediated schema
- Mappings: $S_i \subseteq Q_i(M)$
Local-as-view data integration

- Data sources
- Mediated schema
- Mappings: $S_i \subseteq Q_i(M)$
- Query $Q$
Certain-data case
Certain-data case

- Mediated database (logical)
  - Contains all sources
Certain-data case

- Mediated database (logical)
  - Contains all sources
  - Multiple possibilities
Certain-data case

- Mediated database (logical)
  - Contains all sources
  - Multiple possibilities
Certain-data case

- Mediated database (logical)
  - Contains all sources
  - Multiple possibilities
- Many candidates $Q(M_i)$
Certain-data case

- Mediated database (logical)
  - Contains all sources
  - Multiple possibilities
- Many candidates $Q(M_j)$
  - Use one with minimum information
Mediated database (logical)
- Contains all sources
- Multiple possibilities
Many candidates $Q(M_j)$
- Use one with minimum information
"Certain answers" $= Q(M_3) \subseteq Q(M_j)$ for all $M_j$
Certain-data case

- Mediated database (logical)
  - Contains all sources
  - Multiple possibilities
- Many candidates $Q(M_j)$
  - Use one with minimum information
- “Certain answers” = $Q(M_3) \subseteq Q(M_j)$ for all $M_j$
Scope
Scope

- Not in this talk
  - Multiple tables, non-identity source mapping queries
  - Probabilities
  - Efficient representations
  - Efficient query answering
Scope

- Not in this talk
  - Multiple tables, non-identity source mapping queries
  - Probabilities
  - Efficient representations
  - Efficient query answering

- In this talk
  - Foundations that enable the above
Uncertain database

\[
\begin{array}{c|c}
T & W^1 & W^2 \\
2, A & 2, A & 2, B \\
2, B & & \\
\end{array}
\]
Uncertain database

- Uncertain database $U$
- Tuple set $T$
- Possible worlds $PW = \{ W^1, W^2, \ldots, W^n \}$
- $W^i \subseteq T$
Uncertain database

- Uncertain database $U$
- Tuple set $T$
- Possible worlds $PW = \{ W^1, W^2, \ldots, W^n \}$
- $W^i \subseteq T$
- Suppose $PW = 2^T$
Uncertain database

- Uncertain database $U$
- Tuple set $T$
- Possible worlds $PW = \{W^1, W^2, \ldots, W^n\}$
- $W^i \subseteq T$
- Suppose $PW = 2^T$
- Suppose $T$ also contained $3, P$
- Absence of a tuple is also information
Containment

Single-possible-world databases

$U_1$

$T_1$
Containment

Single-possible-world databases
Containment

Single-possible-world databases

U_1 = (T_1, \{W_1\}) and U_2 = (T_2, \{W_2\})
Containment

Single-possible-world databases

- \( U_1 = (T_1, \{W_1\}) \) and \( U_2 = (T_2, \{W_2\}) \)
- Containment: \( U_1 \subseteq U_2 \)
Containment

Single-possible-world databases

- $U_1 = (T_1, \{W_1\})$ and $U_2 = (T_2, \{W_2\})$
- Containment: $U_1 \subseteq U_2$
- $W_1 \subseteq W_2$
Containment

Single-possible-world databases

- $U_1 = (T_1, \{W_1\})$ and $U_2 = (T_2, \{W_2\})$
- Containment: $U_1 \subseteq U_2$
  - $W_1 \subseteq W_2$ (presence)
  - $(T_1 - W_1) \subseteq (T_2 - W_2)$ (absence)
Containment

General case
Containment

General case

\[ U_1 \subseteq U_2 \] is defined as
**Containment**

**General case**

- \( U_1 \subseteq U_2 \) is defined as
  - For all \( W_2 \) from \( U_2 \)
  - Exists \( W_1 \) from \( U_1 \)
    - \((T_1, \{W_1\}) \subseteq (T_2, \{W_2\})\)
Containment

General case

- $U_1 \subseteq U_2$ is defined as
- For all $W_2$ from $U_2$
- Exists $W_1$ from $U_1$
- $(T_1, \{W_1\}) \subseteq (T_2, \{W_2\})$
Containment

General case

- \( U_1 \subseteq U_2 \) is defined as
  - For all \( W_2 \) from \( U_2 \)
  - Exists \( W_1 \) from \( U_1 \)
  - \((T_1, \{W_1\}) \subseteq (T_2, \{W_2\})\)
Containment

General case

- $U_1 \subseteq U_2$ is defined as
  - For all $W_2$ from $U_2$
  - Exists $W_1$ from $U_1$
  - $(T_1, \{W_1\}) \subseteq (T_2, \{W_2\})$
Containment

General case

- $U_1 \subseteq U_2$ is defined as
  - For all $W_2$ from $U_2$
  - Exists $W_1$ from $U_1$
  - $(T_1, \{W_1\}) \subseteq (T_2, \{W_2\})$
- “Smyth lifting” from power-domains
Containment to integration
Containment to integration

- Mediated database (logical)

![Diagram showing containment to integration]
Containment to integration

- Mediated database (logical)
Containment to integration

- Mediated database (logical)
  - Uncertain
- Query result also uncertain
Containment to integration

- Mediated database (logical)
  - Uncertain
- Query result also uncertain
Containment to integration

- Mediated database (logical)
- Uncertain
- Query result also uncertain
- Strongest correct answer
  - $= Q(M_4) \subseteq Q(M_j)$ for all $M_j$
Less abstract example

- Extract dates of World Series games
Less abstract example

- Extract dates of World Series games

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/26/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
</tbody>
</table>
Less abstract example

- Extract dates of World Series games

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/26/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game1</td>
<td>10/26/09</td>
</tr>
<tr>
<td>Game2</td>
<td>10/26/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
</tr>
</tbody>
</table>
Less abstract example

- Extract dates of World Series games

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Game1</td>
<td>10/26/09</td>
<td></td>
</tr>
<tr>
<td>Game2</td>
<td>10/29/09</td>
<td></td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game1</td>
<td>10/29/09</td>
<td></td>
</tr>
<tr>
<td>Game2</td>
<td>10/31/09</td>
<td></td>
</tr>
<tr>
<td>Game3</td>
<td>10/31/09</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Game1</td>
<td>10/26/09</td>
<td></td>
</tr>
<tr>
<td>Game2</td>
<td>10/29/09</td>
<td></td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game1</td>
<td>10/29/09</td>
<td></td>
</tr>
<tr>
<td>Game2</td>
<td>10/31/09</td>
<td></td>
</tr>
<tr>
<td>Game3</td>
<td>10/31/09</td>
<td></td>
</tr>
</tbody>
</table>
Less abstract example

- Extract dates of World Series games

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/26/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/29/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/31/09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/26/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/31/09</td>
</tr>
</tbody>
</table>
Less abstract example

- Extract dates of World Series games

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/26/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/26/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/29/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/31/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/31/09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Game1</th>
<th>10/26/09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/31/09</td>
</tr>
</tbody>
</table>

Corroboration
Less abstract example

- Extract dates of World Series games

<table>
<thead>
<tr>
<th>Game</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game1</td>
<td>10/26/09</td>
</tr>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game2</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game1</td>
<td>10/26/09</td>
</tr>
<tr>
<td>Game2</td>
<td>10/31/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/31/09</td>
</tr>
<tr>
<td>Game1</td>
<td>10/29/09</td>
</tr>
<tr>
<td>Game2</td>
<td>10/31/09</td>
</tr>
<tr>
<td>Game3</td>
<td>10/31/09</td>
</tr>
</tbody>
</table>

Corroboration

Contradiction
Consistency
Consistency

- Inconsistent
  - A and B are mutually exclusive
  - A and B occur together
Consistency

- Inconsistent
  - A and B are mutually exclusive
  - A and B occur together
  - No uncertain database can contain both
- Consistency definition: a mediated database exists
Consistency

✧ Inconsistent
  ✧ A and B are mutually exclusive
  ✧ A and B occur together
  ✧ No uncertain database can contain both
✧ Consistency definition: a mediated database exists
✧ Queries can’t be answered
Consistency checking
Consistency checking

- First step for query answering
Consistency checking

- First step for query answering
- Given: data sources, mediated schema and mappings
  - Check if consistent
- NP-hard in general
- PTIME
  - Constant number of sources
Consistency checking

- First step for query answering
- Given: data sources, mediated schema and mappings
  - Check if consistent
- NP-hard in general
- PTIME
  - Constant number of sources
  - Sources induce acyclic hypergraph
Another scenario

Actual database
Another scenario

✦ Sources derived from an actual uncertain database
  ✦ Selection query
  ✦ Access control
Another scenario

- Sources derived from an actual uncertain database
  - Selection query
  - Access control
Another scenario

- Sources derived from an actual uncertain database
- Selection query
- Access control
- Goal: best reconstruction of actual database
Another scenario

- Sources derived from an actual uncertain database
  - Selection query
  - Access control
- Goal: best reconstruction of actual database
- Sources cannot have "phantom" worlds
- New containment definition
  - Stricter
Another scenario

- Sources derived from an actual uncertain database
  - Selection query
  - Access control
- Goal: best reconstruction of actual database
- Sources cannot have “phantom” worlds
- New containment definition
  - Stricter

\[
\text{Actual database} \subseteq \notin \text{Source}
\]
Another scenario

- Sources derived from an actual uncertain database
  - Selection query
  - Access control
- Goal: best reconstruction of actual database
- Sources cannot have “phantom” worlds
- New containment definition
  - Stricter
  - “Plotkin lifting”
Beyond this talk
Beyond this talk

- Multiple tables, monotonic queries
Beyond this talk

- Multiple tables, monotonic queries
- Probabilities
Beyond this talk

- Multiple tables, monotonic queries
- Probabilities
  - Deal with inconsistent sources
Beyond this talk

- Multiple tables, monotonic queries
- Probabilities
  - Deal with inconsistent sources
  - Source reliability
Beyond this talk

- Multiple tables, monotonic queries
- Probabilities
  - Deal with inconsistent sources
  - Source reliability
  - Dependent/independent sources
Beyond this talk

- Multiple tables, monotonic queries
- Probabilities
  - Deal with inconsistent sources
  - Source reliability
  - Dependent/independent sources
- Efficient representations and query answering
Beyond this talk

- Multiple tables, monotonic queries
- Probabilities
  - Deal with inconsistent sources
  - Source reliability
  - Dependent/independent sources
- Efficient representations and query answering
- Beyond local-as-view
Uncertain data integration can reduce uncertainty by exploiting redundancy.
Uncertain data integration can reduce uncertainty by exploiting redundancy.

Thank you!