Entity Linkage for Heterogeneous, Uncertain, and Volatile Data

Ekaterini Ioannou

L3S Research Center
Leibniz Universität Hannover

Friday, 15th of April, 2011
Data integration - Entity Linkage

Combine data from various sources and applications
Create a unified view over the data:

- Variations in textual representations
e.g., “J. Web Sem.”, “Journal of Web Semantics”

- Evolving nature of data
e.g., “Jacqueline Lee Bouvier”, “Jackie Kennedy”, “Jackie Onassis”

- Lack of a global coordination for identifier assignment
Data integration - Entity Linkage

Combine data from various sources and applications

Create a unified view over the data:

- Variations in textual representations
  e.g., “J. Web Sem.”, “Journal of Web Semantics”

- Evolving nature of data
  e.g., “Jacqueline Lee Bouvier”, “Jackie Kennedy”, “Jackie Onassis”

- Lack of a global coordination for identifier assignment

Entity Linkage

Identifying data describing the same real world object
Entity Linkage - Existing Approaches

1. Atomic similarity metrics
   compute matching of two entities [CRF03]

2. Similarity of data sets
   deals with entities that are provided as sets [OS99, DH05]

3. Entity inner-relationships
   improves matching through available relationships [KM06, DHM05]

4. Model alternative matches as uncertain data
   processing follows the possible worlds semantics [AFM06]
Typical Process [EIV07]:

1. Detect entity linkages (with probabilities)
2. Merge entities (those above a threshold)
3. Query answering over database with merged entities

Data in modern Web applications is not static
Change syntax, structure, and semantics [Vel08, EIV07]
→ Mechanism for addressing new challenges
**Motivating Example**

<table>
<thead>
<tr>
<th>Entities: set of attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attributes:</strong> name-value pair</td>
</tr>
<tr>
<td>Aligned with dataspaces [HFM06] and idea of concepts [DKP⁺09]</td>
</tr>
</tbody>
</table>

| title: Harry Potter and the Chamber of Secrets | 0.6 |
| stinting: Daniel Radcliffe | 0.7 |
| stinting: Emma Watson | 0.4 |
| writer: J.K. Rowling | 0.6 |
| genre: Fantasy | 0.6 |
| title: Harry Potter and the Chamber of Secrets | 0.8 |
| genre: Fantasy | 0.8 |
| writer: J.K. Rowling | 0.7 |

| name: International Business Machines | 0.9 |
| base: New York | 0.7 |
| date: 2002 | 0.7 |

(existing entities)
### Motivating Example

<table>
<thead>
<tr>
<th>title:</th>
<th>Harry Potter and the Chamber of Secrets</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>starring:</td>
<td>Daniel Radcliffe</td>
<td>0.7</td>
</tr>
<tr>
<td>starring:</td>
<td>Emma Watson</td>
<td>0.4</td>
</tr>
<tr>
<td>writer:</td>
<td>J.K. Rowling</td>
<td>0.6</td>
</tr>
<tr>
<td>genre:</td>
<td>Fantasy</td>
<td>0.6</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>title:</td>
<td>Harry Potter and the Chamber of Secrets</td>
<td>0.8</td>
</tr>
<tr>
<td>genre:</td>
<td>Fantasy</td>
<td>0.8</td>
</tr>
<tr>
<td>writer:</td>
<td>J.K. Rowling</td>
<td>0.7</td>
</tr>
</tbody>
</table>

| name:           | International Business Machines          | 0.9 |
| base:           | New York                                 | 0.7 |
| date:           | 2002                                     | 0.7 |

**Entities:** set of attributes  
**Attributes:** name-value pair  
Aligned with dataspaces [HFM06] and idea of concepts [DKP⁺09]
### Motivating Example

**Entities:** set of attributes  
**Attributes:** name-value pair  
Aligned with dataspaces [HFM06] and idea of concepts [DKP+09]

#### Challenges
- **Heterogeneity:**  
  - absence of uniform schema  
  - variations in representations

<table>
<thead>
<tr>
<th>Existing Entities</th>
<th>New Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>title:</strong> Harry Potter and the Chamber of Secrets</td>
<td><strong>title:</strong> Harry Potter and the Chamber of Secrets</td>
</tr>
<tr>
<td><strong>genre:</strong> Fantasy</td>
<td><strong>genre:</strong> Fantasy</td>
</tr>
<tr>
<td><strong>writer:</strong> J.K. Rowling</td>
<td><strong>writer:</strong> J.K. Rowling</td>
</tr>
<tr>
<td><strong>name:</strong> International Business Machines</td>
<td><strong>name:</strong> The Big Blue</td>
</tr>
<tr>
<td><strong>base:</strong> New York</td>
<td><strong>location:</strong> California</td>
</tr>
<tr>
<td><strong>date:</strong> 2002</td>
<td><strong>date:</strong> 2002</td>
</tr>
</tbody>
</table>

**title:** Harry Potter and the Chamber of Secrets  
**starring:** Daniel Radcliffe  
**starring:** Emma Watson  
**writer:** J.K. Rowling  
**genre:** Fantasy  
**name:** International Business Machines  
**base:** New York  
**date:** 2002  
**codename:** The Big Blue  
**location:** California
Motivating Example

**Entities**: set of attributes

**Attributes**: name-value pair

Aligned with dataspaces [HFM06] and idea of concepts [DKP+09]

**Challenges**

- Heterogeneity
- Uncertainty:
  - extraction confidence
  - reliability of source
  - outdated or inconsistent
  - ...

<table>
<thead>
<tr>
<th>title: Harry Potter and the Chamber of Secrets</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>starring: Daniel Radcliffe</td>
<td>0.7</td>
</tr>
<tr>
<td>starring: Emma Watson</td>
<td>0.4</td>
</tr>
<tr>
<td>writer: J.K. Rowling</td>
<td>0.6</td>
</tr>
<tr>
<td>genre: Fantasy</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title: Harry Potter and the Chamber of Secrets</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>genre: Fantasy</td>
<td>0.8</td>
</tr>
<tr>
<td>writer: J.K. Rowling</td>
<td>0.7</td>
</tr>
</tbody>
</table>

| name: International Business Machines          | 0.9 |
| base: New York                                 | 0.7 |
| date: 2002                                     | 0.7 |

| codename: The Big Blue                         | 0.8 |
| location: California                           | 0.5 |

**Existing entities**

**New entities**
Motivating Example

**Entities**: set of attributes  
**Attributes**: name-value pair  
Aligned with dataspaces [HFM06] and idea of concepts [DKP+09]

**Challenges**  
- Heterogeneity  
- Uncertainty  
- Volatile nature of data:  
  - data reduction, addition, and modification
Motivating Example

<table>
<thead>
<tr>
<th>Title</th>
<th>Harry Potter and the Chamber of Secrets 0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starring</td>
<td>Daniel Radcliffe 0.7</td>
</tr>
<tr>
<td>Starring</td>
<td>Emma Watson 0.4</td>
</tr>
<tr>
<td>Writer</td>
<td>J.K. Rowling 0.6</td>
</tr>
<tr>
<td>Genre</td>
<td>Fantasy 0.6</td>
</tr>
</tbody>
</table>

For initial entities:
- merge 1st-2nd
- replace existing entities

<table>
<thead>
<tr>
<th>Title</th>
<th>Harry Potter and the Chamber of Secrets 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>Fantasy 0.8</td>
</tr>
<tr>
<td>Writer</td>
<td>J.K. Rowling 0.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>International Business Machines 0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>New York 0.7</td>
</tr>
<tr>
<td>Date</td>
<td>2002 0.7</td>
</tr>
</tbody>
</table>

Traditional linkage approach
Motivating Example

Traditional linkage approach

For initial entities:
- merge $1^{st}-2^{nd}$
- replace existing entities
Motivating Example

Traditional linkage approach

For initial entities:
- merge 1\textsuperscript{st}-2\textsuperscript{nd}
- replace existing entities

Options for new entities:
1. also merge 4\textsuperscript{th}
2. no merging
### Motivating Example

**Existing Entities**

<table>
<thead>
<tr>
<th>Title</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry Potter and the Chamber of Secrets</td>
<td>0.6</td>
</tr>
<tr>
<td>Genre</td>
<td>Score</td>
</tr>
<tr>
<td>Fantasy</td>
<td>0.6</td>
</tr>
<tr>
<td>Writer</td>
<td>Score</td>
</tr>
<tr>
<td>J.K. Rowling</td>
<td>0.6</td>
</tr>
</tbody>
</table>

**New Entities**

<table>
<thead>
<tr>
<th>Name</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Business Machines</td>
<td>0.9</td>
</tr>
<tr>
<td>Base</td>
<td>Score</td>
</tr>
<tr>
<td>New York</td>
<td>0.7</td>
</tr>
<tr>
<td>Date</td>
<td>Score</td>
</tr>
<tr>
<td>2002</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**Options for Revisiting Existing Entities**

- Merge 1\textsuperscript{st} - 2\textsuperscript{nd}
- Replace existing entities

**Options for New Entities**

1. Also merge 4\textsuperscript{th}
2. Do not merge

### Traditional Linkage Approach

For initial entities:
- merge 1\textsuperscript{st} - 2\textsuperscript{nd}
- replace existing entities

### Problem

Ignores options that would arise from revisiting any of the previous merging decisions
Summary of Approach

Entity linkage process:

- No a-priory merging of entities
- Maintain linkage information alongside the data
- On-the-fly entity-aware query processing

Main subproblems:

1. Modeling Entities and Linkages
2. Efficient Query Processing
3. Detecting Probabilistic Entity Linkage
Outline

1. Introduction
2. Probabilistic Linkage Database (LinkDB)
3. Query Processing for LinkDB
4. Detecting Probabilistic Entity Linkages
5. Conclusions
Outline

1. Introduction
2. Probabilistic Linkage Database (LinkDB)
3. Query Processing for LinkDB
4. Detecting Probabilistic Entity Linkages
5. Conclusions
Entities & Linkages

Linkages:
- $l_{e_i,e_j}$ when entities refer to the same objects
- probabilities reflect belief of $l_{e_i,e_j}$
Example

Query:
\[
\langle \text{name}=\text{“The Big Blue”}, \text{base}=\text{“New York”} \rangle
\]

Assuming no linkages:
zero results

Accepting linkage \(e_4 \equiv e_5\)
answer: \(\text{merge}(e_4, e_5)\)
Example

Query:
\[
\langle \text{writer} = \text{“J.K. Rowling”}, \quad \text{genre} = \text{“Fantasy”} \rangle
\]

Possible Answers:
e_1, e_3
Example

Query:
\[
\langle \text{writer} = \text{"J.K. Rowling"}, \quad \text{genre} = \text{"Fantasy"} \rangle
\]

Possible Answers:
e_1, e_3
\[
\text{merge}(e_1, e_2), e_3
\]
Example

Query:
\[ \langle \text{writer} = \text{“J.K. Rowling”}, \text{genre} = \text{“Fantasy”} \rangle \]

Possible Answers:
e_1, e_3
merge(e_1, e_2), e_3
merge(e_1, e_3)

date: 2002
starring: Daniel Radcliffe
writer: J.K. Rowling
genre: Fantasy

starring: Emma Watson
writer: J.K. Rowling
genre: Fantasy

name: International Business Machines
base: New York
date: 2002
Example

Query:
\langle \text{writer=“J.K. Rowling”}, \\
\text{genre=“Fantasy”} \rangle

Possible Answers:
e_1, e_3 
merge(e_1, e_2), e_3 
merge(e_1, e_3) 
merge(e_1, e_2, e_3)
Possible Worlds - Example [DS04]

\[ S^p = \begin{array}{|c|c|} \hline A & B \\ \hline s_1 & \text{‘m’} & 0.8 \\ s_2 & \text{‘n’} & 0.5 \\ \hline \end{array} \quad T^p = \begin{array}{|c|c|} \hline C & D \\ \hline t_1 & \text{‘p’} & 0.6 \\ \hline \end{array} \]

\[ pwd(D^p) = \begin{array}{|l|l|} \hline \text{database instance} & \text{probability} \\ \hline D_1 = \{s_1, s_2, t_1\} & 0.24 \\ D_2 = \{s_1, t_1\} & 0.24 \\ D_3 = \{s_2, t_1\} & 0.06 \\ D_4 = \{t_1\} & 0.06 \\ D_5 = \{s_1, s_2\} & 0.16 \\ D_6 = \{s_1\} & 0.16 \\ D_7 = \{s_2\} & 0.04 \\ D_8 = \emptyset & 0.04 \\ \hline \end{array} \]

\[ P(D_3) = (1 - P(s_1)) \times P(s_2) \times P(t_1) = 0.2 \times 0.5 \times 0.6 \]
Possible I-world

Linkage Specification is an accepted subset, e.g., $\mathcal{L}^{sp} = \{ e_{1,2}, e_{4,5} \}$
Some L_{sp} are invalid:

\[ L_{sp} = \{ l_{e_1, e_2}, l_{e_4, e_5} \} \]

Linkage Specification is an accepted subset, e.g., \( L^{sp} \).
Linkage Specification is an accepted subset, e.g., $\mathcal{L}^{sp} = \{ l_{e_1,e_2}, l_{e_4,e_5} \}$

Some $\mathcal{L}^{sp}$ are invalid:
Example for $\mathcal{L} = \{ l_{e_1,e_2}, l_{e_2,e_3}, l_{e_1,e_3} \}$

$\mathcal{L}^{sp} = \{ l_{e_1,e_2}, l_{e_1,e_3} \}$ is invalid — transitivity: $e_1 \equiv e_2 \equiv e_3$ AND $e_2 \neq e_3$
Valid linkage specifications  
⇒ possible l-worlds  
Probabilities on linkages are eliminated  
BUT attribute probabilities are still present  
Generate the possible worlds  
(as performed in probabilistic databases)
The different kinds of probabilistic databases

- Probabilistic linkage database
  - Linkage specification

- Probabilistic database with linkages
  - (a possible l-world)

- Regular database
  - (a possible world)
  - Attribute selection

- Probabilistic database
  - (the core)
  - Core computation
On-the-Fly Query Processing

Given a database and a query:

1. generate all possible \( l \)-worlds
2. identify and ignore invalid \( l \)-worlds
3. compute probability of each \( l \)-world
4. generate all possible worlds (for each \( l \)-world)
5. compute probability of each world
6. identify worlds satisfying query
On-the-Fly Query Processing

Given a database and a query:

1. generate all possible $l$-worlds
2. identify and ignore invalid $l$-worlds
3. compute probability of each $l$-world
4. generate all possible worlds (for each $l$-world)
5. compute probability of each world
6. identify worlds satisfying query

$\mapsto$ Prohibitively Expensive (both space and time)
Summary & Contributions

- Combines aspects of entity linkage and of probabilistic databases
- Generic entity-based representation model for highly heterogeneous, and volatile data
- Supports the simultaneous representation of possible linkages between entities alongside the original data
- Uncertainty not only on the attributes of the entities, but also on their linkages

References


Outline

1. Introduction
2. Probabilistic Linkage Database (LinkDB)
3. Query Processing for LinkDB
4. Detecting Probabilistic Entity Linkages
5. Conclusions
Related Work

Recent approaches on managing probabilistic data:
e.g., Trio [ABS⁺06], MayBMS [AKO07], Suciu et al. [DS04, RDS07]

Majority of existing probabilistic techniques:

- Typically the probabilities per tuple (alternative values)
- Based on independence assumption between data
- Focus on efficient query processing
Related Work

Two approaches are more related: [DHY07, AFM06]

Data Integration with Uncertainty [DHY07]:
- Probabilistic mappings between schema information
- Can become input to LinkDB (as entity linkages)

Clean Answers over Dirty Databases [AFM06]:
- Each tuple is an entity
- Matches between entities are known
- No correlations between entities
Representing & Indexing Factors

- Common approach in probabilistic databases is to partition the data into a series of independent groups [AKO07, DS07, RS08, SD07]

- We follow a similar idea to [SD07], since they manage uncertain data with correlations

\( \mathcal{L} \) is the set of linkages, e.g., \( \{ \ell_{e_1, e_2}, \ell_{e_1, e_3}, \ell_{e_4, e_5} \} \)

Factors are pairwise linked entities, e.g., \( \{\{e_1, e_2, e_3\}, \{e_4, e_5\}\} \)

\( \mathcal{L} \) has many factors: \( \mathcal{L}_{f_1}, \mathcal{L}_{f_2}, \ldots \)

Possible \( \ell \)-worlds created as follows:

\[
plw(\langle \mathcal{E}, \mathcal{L}, p^a, p^l \rangle) = \mathcal{L}^{sp}_{f_1} \times \mathcal{L}^{sp}_{f_2} \times \ldots \times \mathcal{L}^{sp}_{f_n}
\]
\[ \mathcal{L} = \{ l_{e_1,e_2}, l_{e_1,e_3}, l_{e_4,e_5} \} \] has two independent factors:

Factor \( f_1 = \{ e_1,e_2,e_3 \} \) for \( \mathcal{L}_{f_1} = \{ l_{e_1,e_2}, l_{e_1,e_3} \} \)

| \( \mathcal{L}_{f_1}^{sp} \) (1) = \( \{ l_{e_1,e_2}, l_{e_1,e_3} \} \) | 0.9 \times 0.6 = 0.54 |
| \( \mathcal{L}_{f_1}^{sp} \) (2) = \( \{ l_{e_1,e_2} \} \) | 0.9 \times (1-0.6) = 0.36 |
| \( \mathcal{L}_{f_1}^{sp} \) (3) = \( \{ l_{e_1,e_3} \} \) | 0.6 \times (1-0.9) = 0.06 |
| \( \mathcal{L}_{f_1}^{sp} \) (4) = \{ \} | (1-0.9) \times (1-0.6) = 0.04 |

Factor \( f_2 = \{ e_4,e_5 \} \) for \( \mathcal{L}_{f_2} = \{ l_{e_4,e_5} \} \)

\( \mathcal{L}_{f_2}^{sp} \) (1) = \( \{ l_{e_4,e_5} \} \) = 0.8

\( \mathcal{L}_{f_2}^{sp} \) (2) = \{ \} = (1-0.8) = 0.2
**Representing & Indexing Factors - Example**

\[ \mathcal{L} = \{ l_{e_1,e_2}, l_{e_1,e_3}, l_{e_4,e_5} \} \] has two independent factors:

<table>
<thead>
<tr>
<th>Factor ( f_1 = { e_1,e_2,e_3 } ) for ( \mathcal{L}<em>{f_1} = { l</em>{e_1,e_2}, l_{e_1,e_3} } )</th>
<th>Factor ( f_2 = { e_4,e_5 } ) for ( \mathcal{L}<em>{f_2} = { l</em>{e_4,e_5} } )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{L}<em>{sp}^{f_1}(1) = { l</em>{e_1,e_2}, l_{e_1,e_3} } )</td>
<td>( \mathcal{L}<em>{sp}^{f_2}(1) = { l</em>{e_4,e_5} } )</td>
</tr>
<tr>
<td>0.9 \times 0.6 = 0.54</td>
<td>0.8</td>
</tr>
<tr>
<td>( \mathcal{L}<em>{sp}^{f_1}(2) = { l</em>{e_1,e_2} } )</td>
<td>( \mathcal{L}_{sp}^{f_2}(2) = { } )</td>
</tr>
<tr>
<td>0.9 \times (1-0.6) = 0.36</td>
<td>(1-0.8) = 0.2</td>
</tr>
<tr>
<td>( \mathcal{L}<em>{sp}^{f_1}(3) = { l</em>{e_1,e_3} } )</td>
<td></td>
</tr>
<tr>
<td>0.6 \times (1-0.9) = 0.06</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{L}_{sp}^{f_1}(4) = { } )</td>
<td></td>
</tr>
<tr>
<td>(1-0.9) \times (1-0.6) = 0.04</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Possible ( l )-world</th>
<th>Required Merges</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l_1 = { l_{e_1,e_2}, l_{e_1,e_3}, l_{e_4,e_5} } )</td>
<td>( e_1 \equiv e_2 \equiv e_3, e_4 \equiv e_5 )</td>
<td>0.54 \times 0.8 = 0.432</td>
</tr>
<tr>
<td>( l_2 = { l_{e_1,e_2}, l_{e_1,e_3} } )</td>
<td>( e_1 \equiv e_2 \equiv e_3, e_4, e_5 )</td>
<td>0.54 \times 0.2 = 0.108</td>
</tr>
<tr>
<td>( l_3 = { l_{e_1,e_2}, l_{e_4,e_5} } )</td>
<td>( e_1 \equiv e_2, e_3, e_4 \equiv e_5 )</td>
<td>0.36 \times 0.8 = 0.288</td>
</tr>
<tr>
<td>( l_4 = { l_{e_1,e_2} } )</td>
<td>( e_1 \equiv e_2, e_3, e_4, e_5 )</td>
<td>0.36 \times 0.2 = 0.072</td>
</tr>
<tr>
<td>( l_5 = { l_{e_1,e_3}, l_{e_4,e_5} } )</td>
<td>( e_1 \equiv e_3, e_2, e_4 \equiv e_5 )</td>
<td>0.06 \times 0.8 = 0.048</td>
</tr>
<tr>
<td>( l_6 = { l_{e_1,e_3} } )</td>
<td>( e_2, e_1 \equiv e_3, e_4, e_5 )</td>
<td>0.06 \times 0.2 = 0.012</td>
</tr>
<tr>
<td>( l_7 = { l_{e_4,e_5} } )</td>
<td>( e_1, e_2, e_3, e_4 \equiv e_5 )</td>
<td>0.04 \times 0.8 = 0.032</td>
</tr>
<tr>
<td>( l_8 = { } )</td>
<td>( e_1, e_2, e_3, e_4, e_5 )</td>
<td>0.04 \times 0.2 = 0.008</td>
</tr>
</tbody>
</table>
Deciding the Entity Merges

Exploit factors to avoid considering all the possible l-worlds:

1. For each query condition we create an entity set $E_i$ with the entities satisfying the specific attribute

2. Cartesian product of these sets with the condition that the entities are of the same factor
Deciding the Entity Merges

Exploit factors to avoid considering all the possible l-worlds:

1. For each query condition we create an entity set $E_i$ with the entities satisfying the specific attribute

2. Cartesian product of these sets with the condition that the entities are of the same factor

Example
$Q$: \(<\text{starring} = \text{“Emma Watson”}, \text{date} = \text{“2002”}\>\)
Deciding the Entity Merges

Exploit factors to avoid considering all the possible l-worlds:

1. For each query condition we create an entity set $E_i$ with the entities satisfying the specific attribute
2. Cartesian product of these sets with the condition that the entities are of the same factor

**Example**

$Q$: $\langle$ starring="Emma Watson", date="2002"$\rangle$

1st Condition: $e_1, e_2 \rightarrow E_1 = \{f_1 - e_1, f_1 - e_2\}$

2nd Condition: $e_2, e_5 \rightarrow E_2 = \{f_1 - e_2, f_2 - e_5\}$
Deciding the Entity Merges

Exploit factors to avoid considering all the possible l-worlds:

1. For each query condition we create an entity set $E_i$ with the entities satisfying the specific attribute

2. Cartesian product of these sets with the condition that the entities are of the same factor

Example
$Q$: ⟨starring="Emma Watson", date="2002"⟩

1st Condition: $e_1, e_2 \rightarrow E_1=\{f_1-e_1, f_1-e_2\}$

2nd Condition: $e_2, e_5 \rightarrow E_2=\{f_1-e_2, f_2-e_5\}$

Cartesian product: $\langle f_1-e_1, f_1-e_2 \rangle$ and $\langle f_1-e_2, f_1-e_2 \rangle$
Deciding the Entity Merges

Exploit factors to avoid considering all the possible l-worlds:

1. For each query condition we create an entity set $E_i$ with the entities satisfying the specific attribute

2. Cartesian product of these sets with the condition that the entities are of the same factor

Example

$Q$: $\langle$ starring=“Emma Watson”, date=“2002” $\rangle$

1st Condition: $e_1, e_2 \rightarrow E_1=\{f_1-e_1, f_1-e_2\}$

2nd Condition: $e_2, e_5 \rightarrow E_2=\{f_1-e_2, f_2-e_5\}$

Cartesian product: $\langle f_1-e_1, f_1-e_2 \rangle$ and $\langle f_1-e_2, f_1-e_2 \rangle$

$\mapsto$ merge($e_1, e_2$), and merge($e_2$)
Computing l-world probabilities

Probability given a query:

\[ \prod_{i=1}^{m} Pr(L_{f_i}^{sp} \mid c_m) \], where \( c_m \) are the conditions describing a merge

- To reduce computation time we consider only the maximum probability
- Create a weighted undirected graph \( G \):
  - nodes are the entities from linkages \( l_{e_i,e_j} \)
  - edges are the linkages \( l_{e_i,e_j} \)
- \( merge(e_1,e_2,\ldots,e_n) \) is a spanning tree connecting \( e_1,e_2,\ldots,e_n \)
- Algorithm is finding shortest paths in graphs
Probabilities of the attributes, specifically in the case of duplication
Dependencies that may exist among attributes

A. Independent Attributes
- No restrictions, i.e., no correlations between attributes
- An entity generated for each merge
- \( \text{merge}(e_1, \ldots, e_n) = \langle \text{id}', \cup_{i=1}^n e_i.A \rangle \)

B. Exclusive Attributes
- An entity must have at most one occurrence of such attributes
- Cluster exclusive attributes, i.e., \( M = \{ e_1.\alpha_i, e_1.\alpha_j, \ldots \} \)
- \( \text{merge}(e_1, \ldots, e_n) = \langle \text{id}', A \rangle \), where
  \[ A \subseteq (M_1 \times M_2 \times \ldots \times M_m) \cup \{ \alpha | \alpha \notin \cup_{i=1}^m M_i.\alpha \} \]
Possible worlds and their probabilities - Example

Consider exclusive attributes (name-value pair):

- starring: “Daniel Radcliffe”
- starring: “Emma Watson”

Figure shows the possible worlds for $\text{merge}(e_1,e_2)$

<table>
<thead>
<tr>
<th>$\text{aid.}$</th>
<th>name</th>
<th>value</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bullet a_{10}$</td>
<td>starring</td>
<td>Daniel Radcliffe</td>
<td>0.7</td>
</tr>
<tr>
<td>$\diamond a_{11}$</td>
<td>starring</td>
<td>Emma Watson</td>
<td>0.4</td>
</tr>
<tr>
<td>$a_{12}$</td>
<td>writer</td>
<td>J.K. Rowling</td>
<td>0.6</td>
</tr>
<tr>
<td>$a_{13}$</td>
<td>genre</td>
<td>Fantasy</td>
<td>0.6</td>
</tr>
<tr>
<td>$\bullet a_{20}$</td>
<td>starring</td>
<td>Daniel Radcliffe</td>
<td>0.5</td>
</tr>
<tr>
<td>$\diamond a_{21}$</td>
<td>starring</td>
<td>Emma Watson</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Consider exclusive attributes (name-value pair):

- starring: “Daniel Radcliffe”
- starring: “Emma Watson”

Figure shows the possible worlds for $\text{merge}(e_1, e_2)$

<table>
<thead>
<tr>
<th>aid.</th>
<th>name</th>
<th>value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{10}$</td>
<td>starring</td>
<td>Daniel Radcliffe</td>
<td>0.7</td>
</tr>
<tr>
<td>$a_{11}$</td>
<td>starring</td>
<td>Emma Watson</td>
<td>0.4</td>
</tr>
<tr>
<td>$a_{12}$</td>
<td>writer</td>
<td>J.K. Rowling</td>
<td>0.6</td>
</tr>
<tr>
<td>$a_{13}$</td>
<td>genre</td>
<td>Fantasy</td>
<td>0.6</td>
</tr>
<tr>
<td>$a_{20}$</td>
<td>starring</td>
<td>Daniel Radcliffe</td>
<td>0.5</td>
</tr>
<tr>
<td>$a_{21}$</td>
<td>starring</td>
<td>Emma Watson</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Experimental Evaluation - Influence of Linkages

Movie Dataset:

- 13,435 movies (23,182 IMDb, & 28,040 DBpedia)

- Two string similarity methods: Jaccard and Jaro

![Graph showing the influence of linkages with precision and recall metrics for Jaccard and Jaro methods.]
Experimental Evaluation - Influence of Linkages

Movie Dataset:

- 13,435 movies (23,182 IMDb, & 28,040 DBpedia)
- Two string similarity methods: Jaccard and Jaro

- Few factors have a large size
- Less overall processing time
Experimental Evaluation

Algorithms:

- **EAQP**: our approach for entity-aware query processing
- **ELA**: entity linkage techniques with unmerged results [WMK+09]
- **PDBA**: probabilistic databases (only for efficiency) [AFM06]

Cora Dataset:

- Probabilistic entity linkages for publication authors
- 9,774 author descriptions that refer to 2,882 real world objects

<table>
<thead>
<tr>
<th>Entity Linkages (under threshold $t$)</th>
<th>$t=0.52$</th>
<th>$t=0.58$</th>
<th>$t=0.62$</th>
<th>$t=0.68$</th>
<th>$t=0.72$</th>
<th>$t=0.78$</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,440</td>
<td>12,012</td>
<td>10,775</td>
<td>6,394</td>
<td>5,985</td>
<td>4,184</td>
<td></td>
</tr>
</tbody>
</table>

Experimental Evaluation

Effectiveness:

- F-measure: weighted harmonic mean of precision/recall
- EAQP exhibits a higher F-measure than ELA
- Higher difference for threshold values 0.65-0.75
Experimental Evaluation

Effectiveness:

- F-measure: weighted harmonic mean of precision/recall
- EAQP exhibits a higher F-measure than ELA
- Higher difference for threshold values 0.65-0.75

Efficiency:

- Small increase in time
- Remains under 70 msec.
- Scalable methodology
Summary & Contributions

- Methodology to efficiently compute the answers for entity queries under probabilistic linkages
- Additional valid query answering results, compared to those of entity linkage and probabilistic databases
- Reasoning about the entity linkages is on the fly, i.e., results inferred by query conditions


Outline

1. Introduction
2. Probabilistic Linkage Database (LinkDB)
3. Query Processing for LinkDB
4. Detecting Probabilistic Entity Linkages
5. Conclusions
Related Work

Existing approaches:

- Off-line processing and merging of the entities [EIV07]
- Few approaches showed that relationships improve effectiveness, e.g., [DHM05, KM06]
- Improvements through relationships and propagation of matching results

Probabilistic Entity Linkages:

- Incremental computation
- Easier adaptation when new data is available
Bayesian Networks - Overview

Probabilistic graphical models for reasoning under uncertainty

- **Nodes**: variables with two or more possible states
- **Edges**: cause-effect (observed) relationships

Nodes are accompanied with:

- Unconditional probability (no parents)
- Conditional probability (given parents)

Probabilistic Inference:
Determines (given any new effects) the conditional probabilities of cause nodes
Bayesian Networks - Overview

Diagram of Bayesian Networks:
- **City** with states: Interesting 35%, Boring 15%
- **Opera** with states: Yes 100%, No 0%
- **Nice Cuisine** with states: Yes 0%, No 100%
- **Museum** with states: Three 100%, Four 0%

Diagram of another set of Bayesian Networks:
- **City** with states: Interesting 63%, Boring 37%
- **Opera** with states: Yes 100%, No 0%
- **Nice Cuisine** with states: Yes 0%, No 100%
- **Museum** with states: Three 0%, Four 100%
Structure of the Bayesian Network

Nodes in the Bayesian network:

- **Linkage**: possible match between entities
- **Supporting evidence**: observed similarities (Soundex, StringSim)
- **Direct-Relation**: related resources
- **Deductive-Relation**: indirect related resources
Structure of the Bayesian Network

Nodes in the Bayesian network:

- **Linkage**: possible match between entities
- **Supporting evidence**: observed similarities (Soundex, StringSim)
- **Direct-Relation**: related resources
- **Deductive-Relation**: indirect related resources

Cause-effect relationships in the Bayesian network:

<table>
<thead>
<tr>
<th>Effect Nodes</th>
<th>(1) Evidence</th>
<th>(2) Direct-Rel.</th>
<th>(3) Deductive-Rel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cause nodes</td>
<td>(1) Linkage</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>(2) Ded.-Rel.</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Incremental Computation of the Network

Step 1 - Add Evidence/Entity nodes

- Compare new with existing entities
- Generate possible matches, i.e., entity linkages
  \[ P[ e_{paper77.author1} = e_{paper127.author1} ] \]
- Add entity/evidence nodes
- Set state of evidence nodes (observed effect)
Incremental Computation of the Network

Step 1 - Add Evidence/Entity nodes

Step 2 - Add Direct-Relation nodes

- Add dir-rel node and cause-effect relationships

$$P \left[ e_{\text{paper}77.\text{author}1} = e_{\text{paper}127.\text{author}1} \right]$$

$$\mapsto \text{dir-rel}(e_{\text{paper}77}, e_{\text{paper}127})$$
Incremental Computation of the Network

Step 1 - Add Evidence/Entity nodes
Step 2 - Add Direct-Relation nodes
Step 3 - Add Deductive-Relation nodes

- Transitive relations:
  \[
  \text{dir-rel}(e_{\text{paper}77}, e_{\text{paper}127}) \quad \text{dir-rel}(e_{\text{paper}127}, e_{\text{email}128})
  \]

- Add ded-rel node and cause-effect relationships
- Stop mechanism using evidence density
Incremental Computation of the Network

Step 1 - Add Evidence/Entity nodes
Step 2 - Add Direct-Relation nodes
Step 3 - Add Deductive-Relation nodes
Step 4 - Update the Linkages
   • Probabilistic Inference
   • Update the entity linkages in the dataset
Example

When $R$ is activated:
- Receives messages from parent and children nodes
- Computes its own belief
- Sends messages to parent and children nodes
Dataset & Methodology

Collection of publications from CiteSeer

Name variants:
- Example $\mapsto$ “J. Antonisse” ; “Antonisse, H. J.” ; “Antonisse”
- Maximum is 88 different entities for the same object

Dataset Information:
- 1563 publications
- 2882 triples describing authors
- 9774 matches between authors
Precision & Recall

- Incremental addition of publications
- Evaluation of linkages for different probability thresholds
- Maintain precision and recall values for the different probability thresholds
Summary & Contributions

- Modeling the entity linkage problem as a Bayesian network
- No need to reprocess data for recomputing linkages, as performed in traditional approaches
- Incremental update of linkages when new information arrives
- Evaluation illustrates efficiency and effectiveness of approach


Outline

1. Introduction
2. Probabilistic Linkage Database (LinkDB)
3. Query Processing for LinkDB
4. Detecting Probabilistic Entity Linkages
5. Conclusions
Conclusions

- Entity linkage methodology focusing on heterogeneous, uncertain, and volatile data
- Generic data model for entities and linkages between entities
- The model is probabilistic, with attribute and linkage uncertainty
- Entity-based query mechanism that exploits linkage information and uncertainty for retrieving entities
- Detection and generation of probabilistic entity linkages
Future Work

Incremental and Adaptive Entity Linkage Index

- Processing based on the popularity of entities
- Frequently requested entities: maintain linkages and merges
- Rarely requested entities: no need to process them

Scaling Entity Linkage to Large Collections

- Investigating blocking techniques, i.e., separating the data into blocks and comparing only the data inside each block
- Existing approaches rely on the homogeneity
- Need of mechanisms for building blocks, scheduling block processing, deciding when to stop processing
<table>
<thead>
<tr>
<th>Reference</th>
<th>Authors</th>
<th>Title</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>[AKO07]</td>
<td>Lyublena Antova, Christoph Koch, and Dan Olteanu.</td>
<td>10^{10^6} worlds and beyond: Efficient representation and processing of incomplete information.</td>
<td>ICDE, 2007.</td>
</tr>
<tr>
<td>Introduction</td>
<td>LinkDB</td>
<td>Query Processing</td>
<td>Detecting Linkages</td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Ref.</th>
<th>Authors</th>
<th>Title</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>[RS08]</td>
<td>Christopher Re and Dan Suciu.</td>
<td>Managing probabilistic data with mystiq: The can-do, the could-do, and the can’t-do.</td>
<td>SUM, 2008.</td>
</tr>
<tr>
<td>[WMK\textsuperscript{+}09]</td>
<td>Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina.</td>
<td>Entity resolution with iterative blocking.</td>
<td>SIGMOD Conference, 2009.</td>
</tr>
</tbody>
</table>
When $R$ is activated:

Receives $\lambda_{r_1}(R)$, $\lambda_{r_2}(R)$, $\pi_{R}(l_1)$, $\pi_{R}(l_2)$

Computes $BEL(R) = \alpha \lambda(R) \pi(R)$, where

$$\lambda(R) = \lambda_{r_1}(R) \lambda_{r_2}(R)$$

$$\pi(R) = P(R|l_1, l_2) \pi_{R}(l_1) \pi_{R}(l_2)$$

Sends message to parent nodes:

$$\lambda_{R}(l_1) = P(R|l_1, l_2) \pi_{R}(l_1) \lambda(R)$$

Sends messages to children nodes:

$$\pi_{r_2}(R) = \pi(R) \lambda_{r_1}(R)$$
Experimental Evaluation

Improvements over ELA:

- Less failures, i.e., empty result sets
- Entities with higher confidence

![Graph showing improvements over ELA](image-url)
Example

- Model the problem using a Bayesian Network
- Based on a collection of matching evidences

```
metadata for publ. #77 (M(r77))={...,
<file:///P77, type, publication>,
<file:///P77, title, ... >,
<file:///P77/a1, name, K. Marriott>,
<file:///P77/a2, name, P. J. Stuckey>}
```

```
metadata for publ. #127 (M(r127))={...,
<file:///P127/a1, name, `Marriott, K'>,
<file:///P127/a2, name, `Sndergaard, H' >,
<file:///P127/a3, name, `Kelly, A'>}
```

```
metadata for email #128 (M(r128))={...,
<file:///E128/to1, name, Kelly A. >,
<file:///E128/from, name, Sndergaard H. >,
<file:///E128/to2, name, Stuckey P. >}
```
Example II

Inference Evaluation

Example II

Example

\[ e_{P77} \]
- type: publication
- title: ...
- \( e_{P77.a1} \)
- \( e_{P77.a2} \)

\[ e_{P77.a1} \]
- name: K. Marriott

\[ e_{P77.a2} \]
- name: P. J. Stuckey

\[ e_{P127} \]
- title: ...
- \( e_{P127.a1} \)
- \( e_{P127.a2} \)
- \( e_{P127.a3} \)

\[ e_{P127.a1} \]
- name: ‘Marriott, K’

\[ e_{E128} \]
- subject: ...
- \( e_{E128.to1} \)
- \( e_{E128.from} \)
- \( e_{E128.to2} \)

\[ e_{E128.to2} \]
- name: Stuckey P.

\[ e_{E128.from} \]
- name: Sndergaard H.