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
$\mapsto$ 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]

Entity Linkage - Existing Approaches

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]

$\mapsto$ Mechanism for addressing new challenges

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

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

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

title:	Harry Potter and the Chamber of Secrets	0.7
date:	2002	0.8
starring:	Daniel Radcliffe	0.5
starring:	Emma Watson	0.9

codename:	The Big Blue	0.8
location:	California	0.5

new entities

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

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

title:	Harry Potter and the Chamber of Secrets	0.7
date:	2002	0.8
starring:	Daniel Radcliffe	0.5
starring:	Emma Watson	0.9

codename:	The Big Blue	0.8
location:	California	0.5

new entities

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

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

title:	Harry Potter and the Chamber of Secrets	0.7
date:	2002	0.8
starring:	Daniel Radcliffe	0.5
starring:	Emma Watson	0.9

codename:	The Big Blue	0.8
location:	California	0.5

new entities

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
  - ...

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

title:	Harry Potter and the Chamber of Secrets	0.7
date:	2002	0.8
starring:	Daniel Radcliffe	0.5
starring:	Emma Watson	0.9

codename:	The Big Blue	0.8
location:	California	0.5

new entities

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

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

Traditional linkage approach

For initial entities:

- merge 1^{st}-2^{nd}
- replace existing entities

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

Traditional linkage approach

For initial entities:

- merge 1^{st}-2^{nd}
- replace existing entities

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

title:	Harry Potter and the Chamber of Secrets	0.7
date:	2002	0.8
starring:	Daniel Radcliffe	0.5
starring:	Emma Watson	0.9

codename:	The Big Blue	0.8
location:	California	0.5

new entities

Traditional linkage approach

For initial entities:

- merge 1^{st}-2^{nd}
- replace existing entities

Options for new entities:

$\xrightarrow{1}$ also merge 4^{th}

$\xrightarrow{2}$ no merging

Motivating Example

title:	Harry Potter and the Chamber of Secrets	0.6
starring:	Daniel Radcliffe	0.7
starring:	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

title:	Harry Potter and the Chamber of Secrets	0.7
date:	2002	0.8
starring:	Daniel Radcliffe	0.5
starring:	Emma Watson	0.9

codename:	The Big Blue	0.8
location:	California	0.5

new entities

Traditional linkage approach

For initial entities:

- merge 1^{st}-2^{nd}
- replace existing entities

Options for new entities:

$\xrightarrow{1}$ also merge 4^{th}

$\xrightarrow{2}$ no merging

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

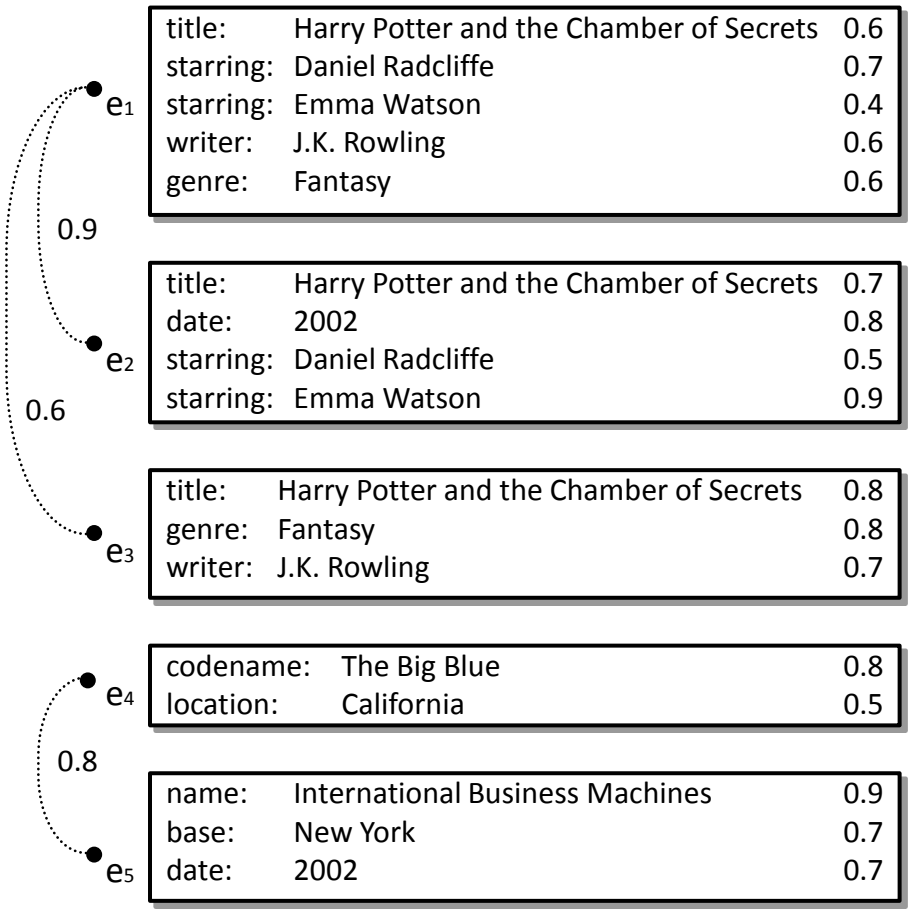
Main subproblems:

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

Outline

Outline

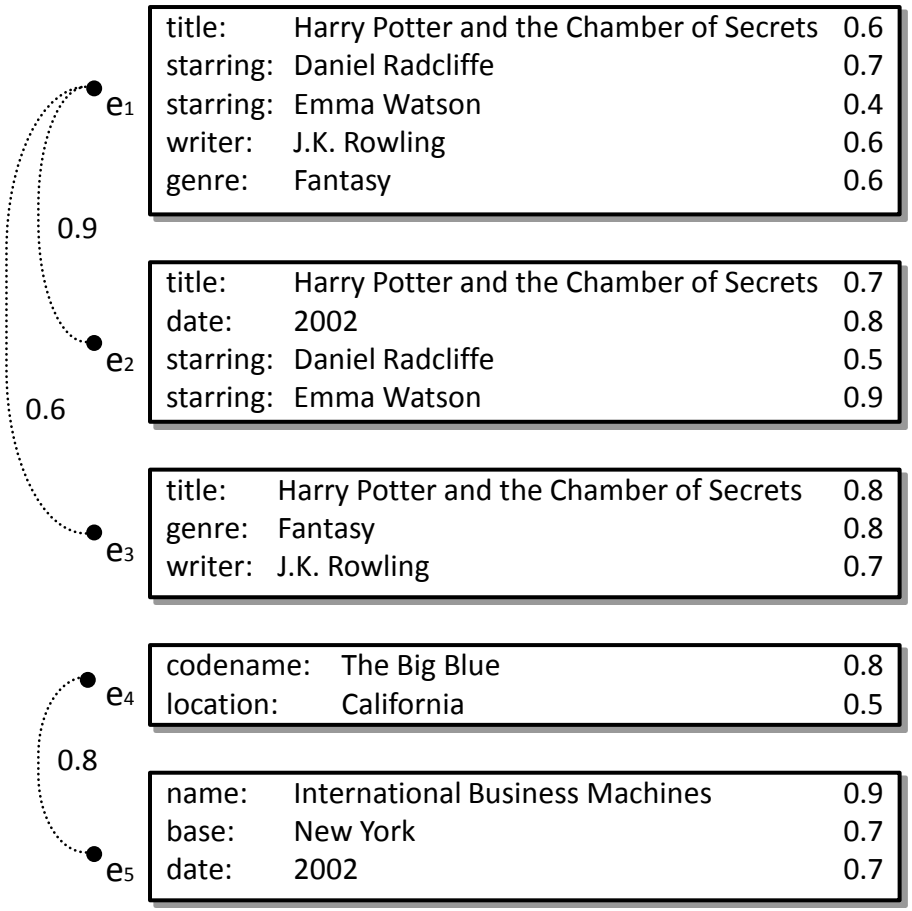
Entities & Linkages



Linkages:

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

Example

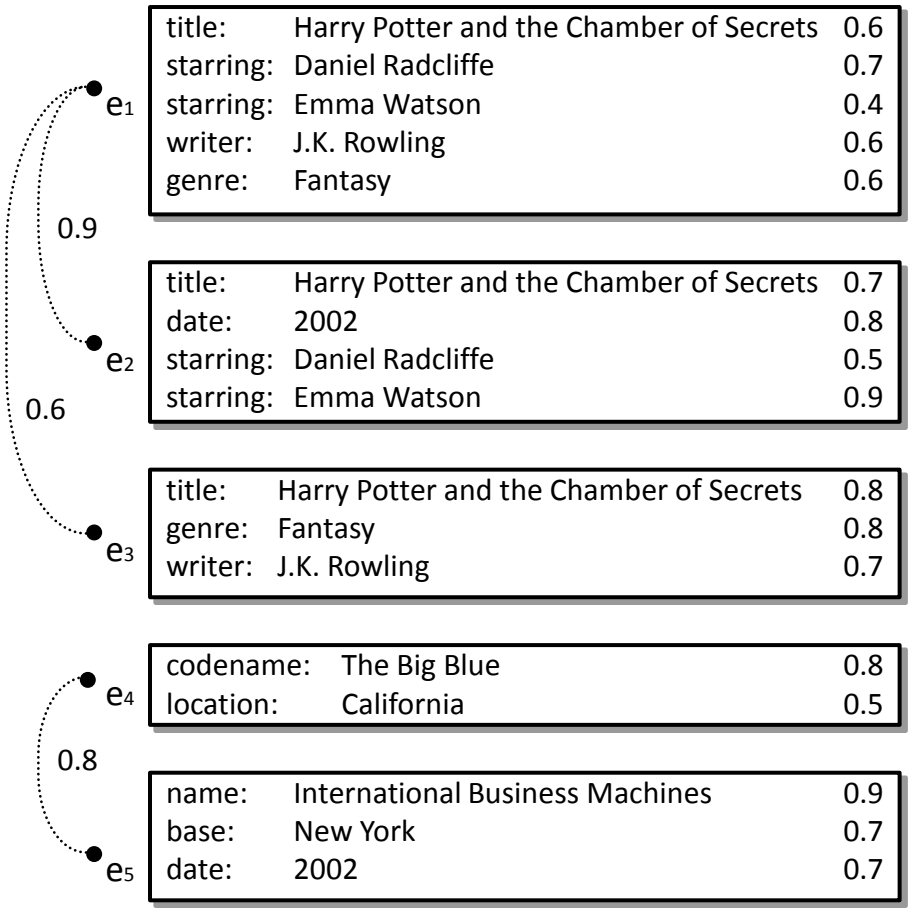


Query:
⟨name="The Big Blue",
base="New York"⟩

Assuming no linkages:
zero results

Accepting linkage $e_4 \equiv e_5$
answer: merge(e_4,e_5)

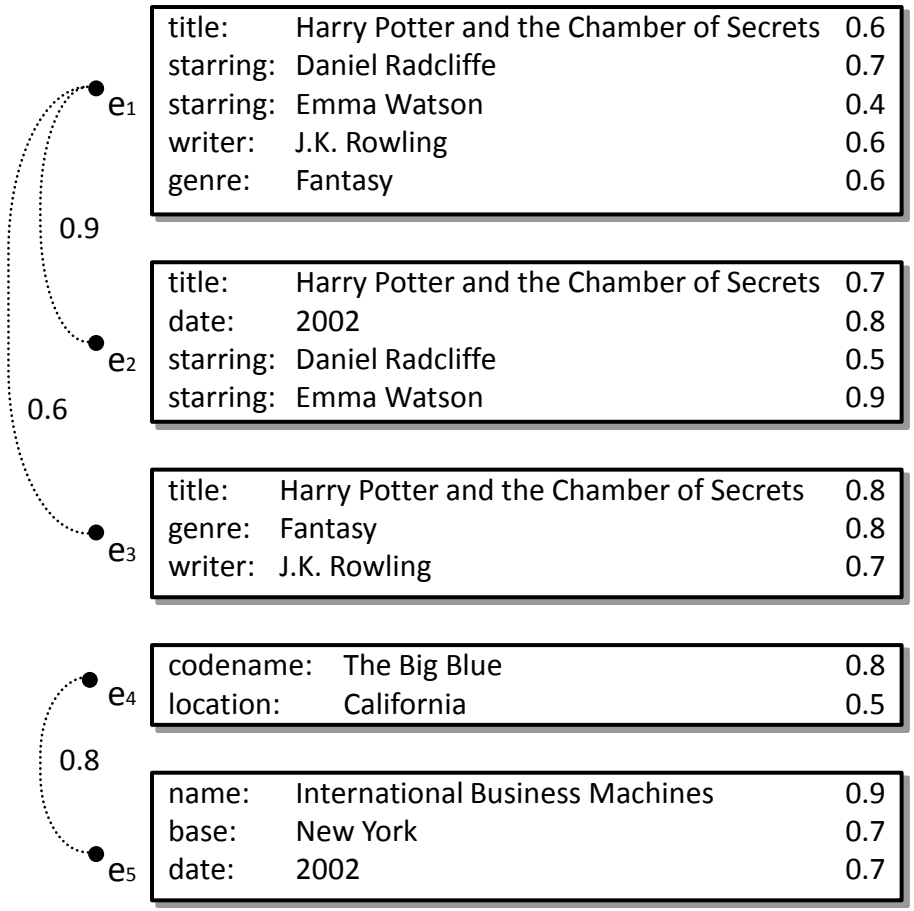
Example



Query:
⟨writer="J.K. Rowling",
genre="Fantasy"⟩

Possible Answers:
e_1, e_3

Example



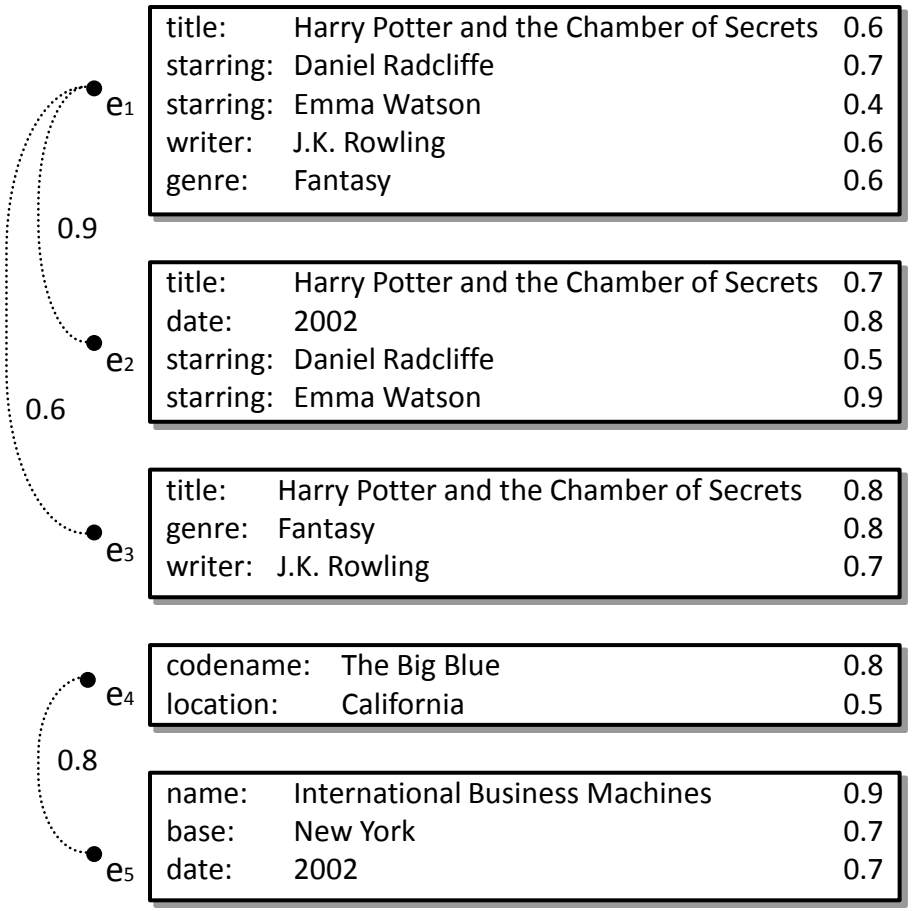
Query:
⟨writer="J.K. Rowling",
genre="Fantasy"⟩

Possible Answers:

e_1, e_3

merge(e_1,e_2), e_3

Example



Query:
⟨writer="J.K. Rowling",
genre="Fantasy"⟩

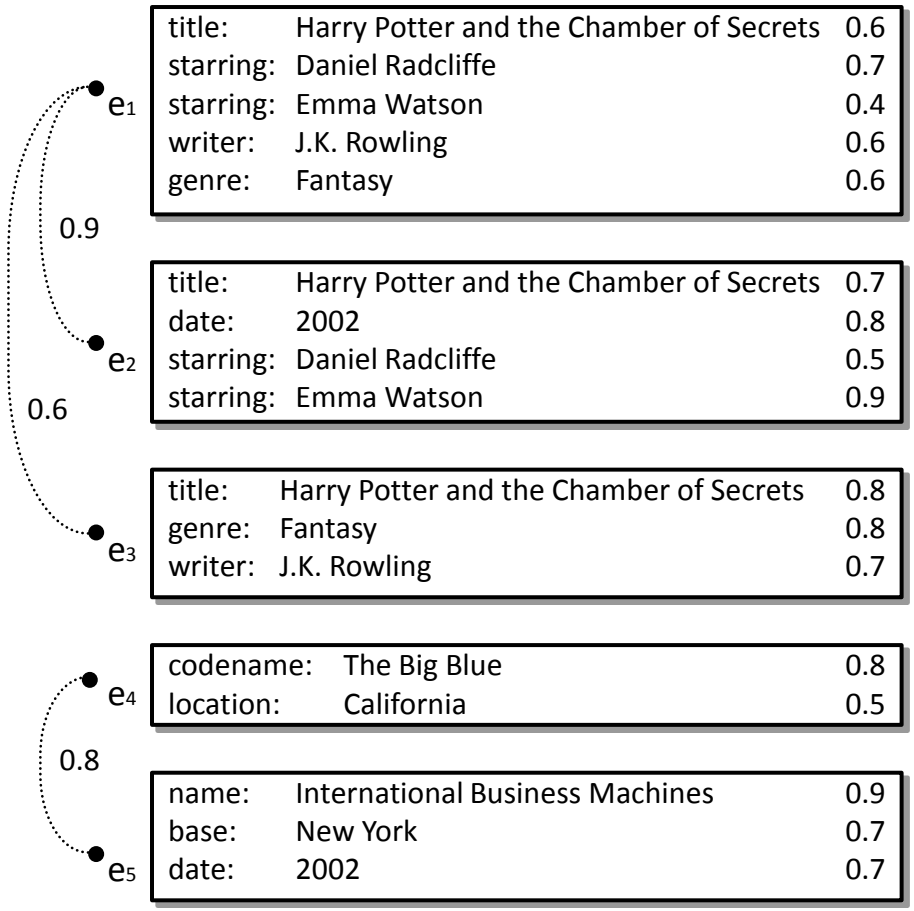
Possible Answers:

e_1, e_3

merge(e_1,e_2), e_3

merge(e_1,e_3)

Example



Query:
⟨writer="J.K. Rowling",
genre="Fantasy"⟩

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]

$S^p =$

	A	B	
s_1	'm'	1	0.8
s_2	'n'	1	0.5

$T^p =$

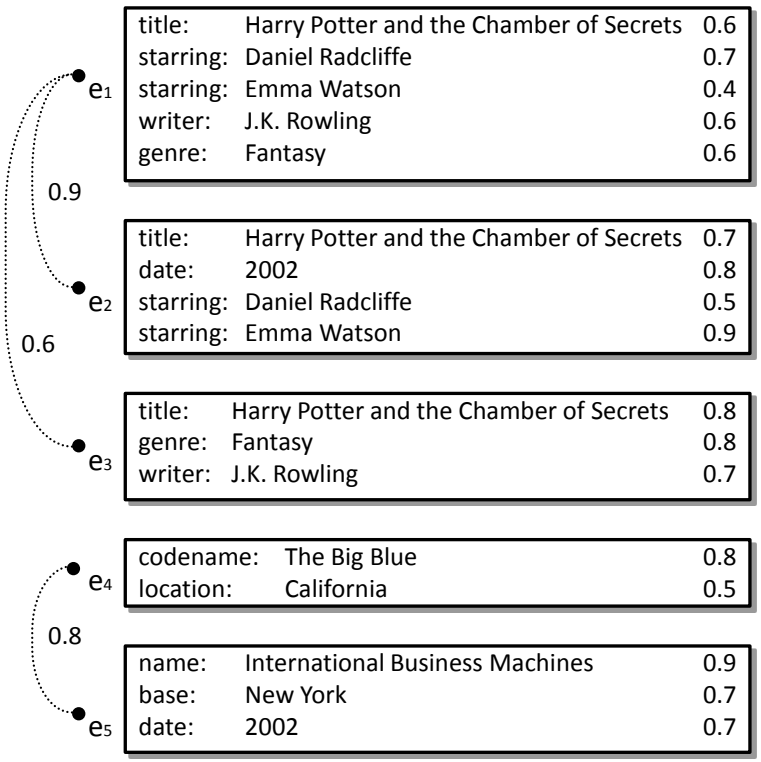
	C	D	
t_1	1	'p'	0.6

$pwd(D^p) =$

database instance	probability
$D_1 = \{s_1, s_2, t_1\}$	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
$D_8 = \phi$	0.04

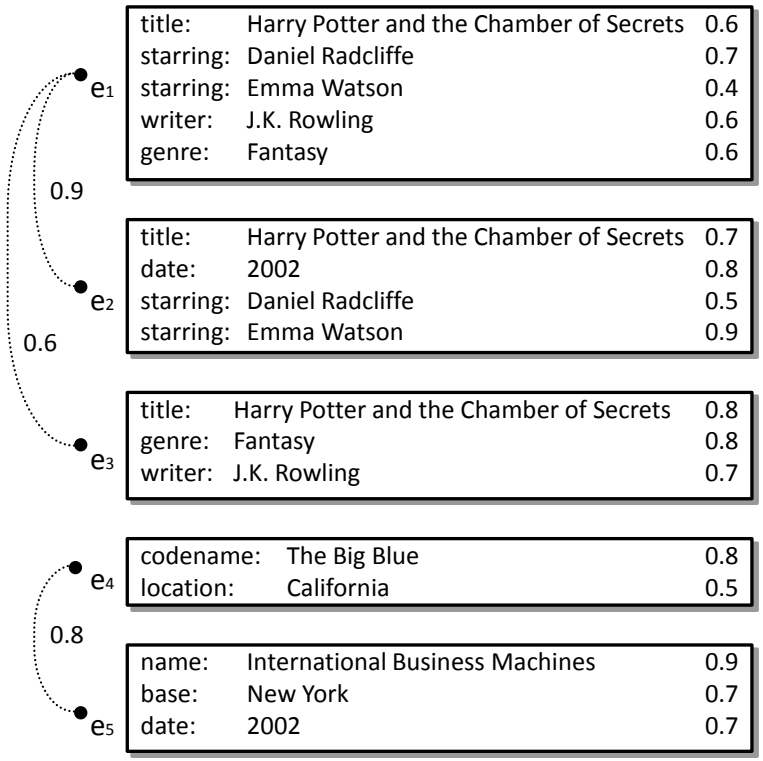
$P(D_3) = (1\text{-}P(s_1)) \times P(s_2) \times P(t_1) = 0.2 \times 0.5 \times 0.6$

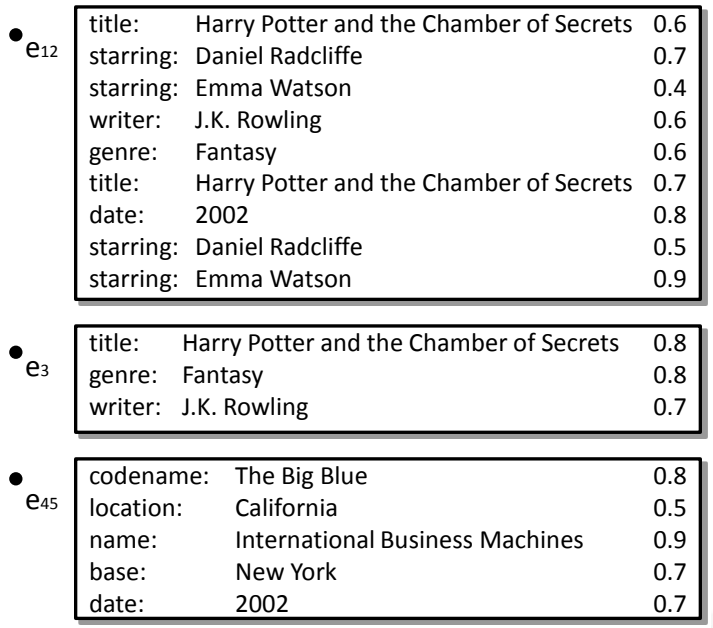
Possible I-world



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

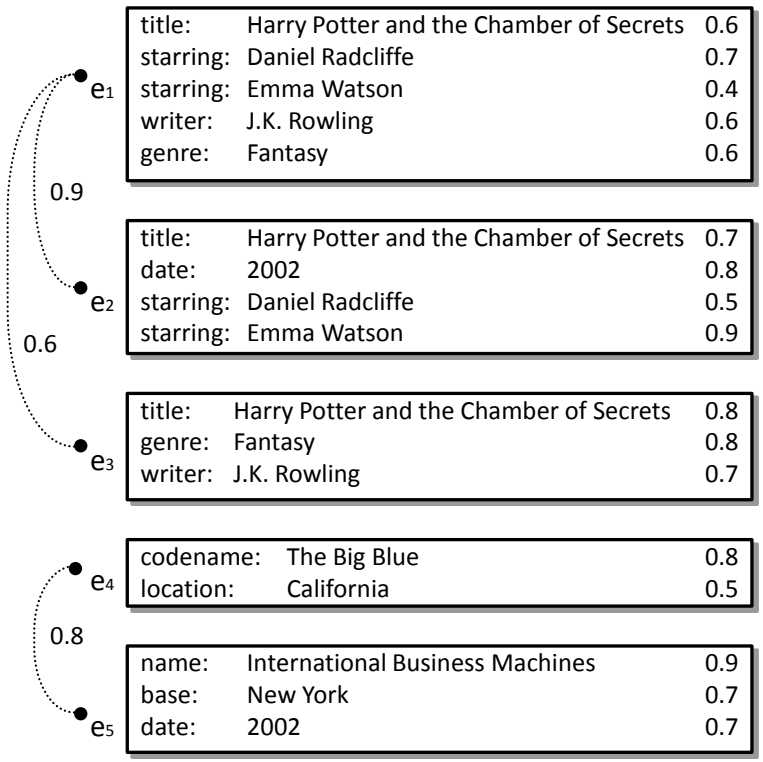
Possible I-world

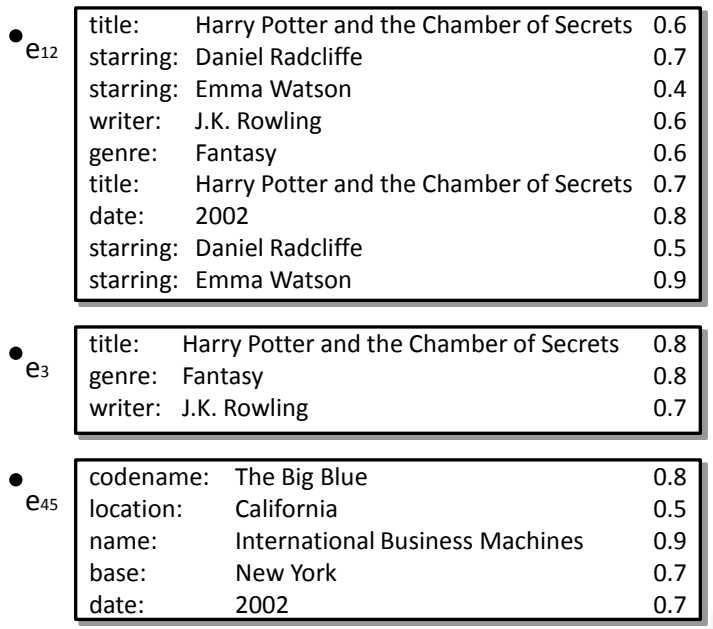




Linkage Specification is an accepted subset, e.g., $\mathcal{L}^{sp} = \{ l_{e_1,e_2}, l_{e_4,e_5} \}$

Possible I-world



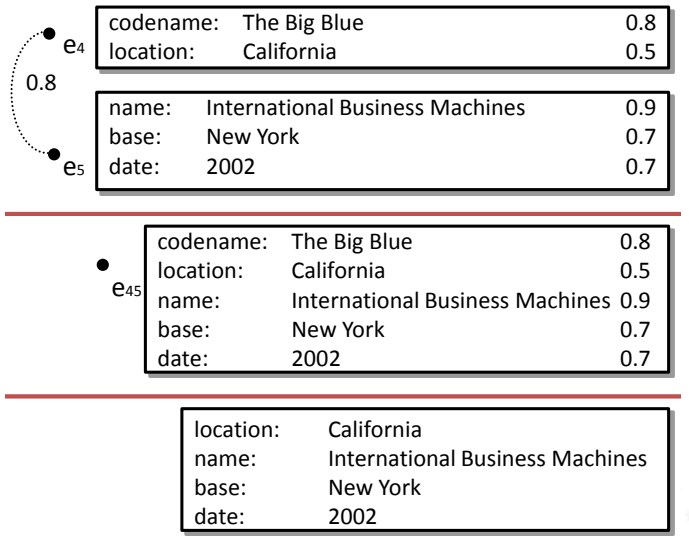


Linkage Specification is an accepted subset, e.g., $\mathcal{L}^{sp}=\{l_{e_1,e_2},\ l_{e_4,e_5}\}$

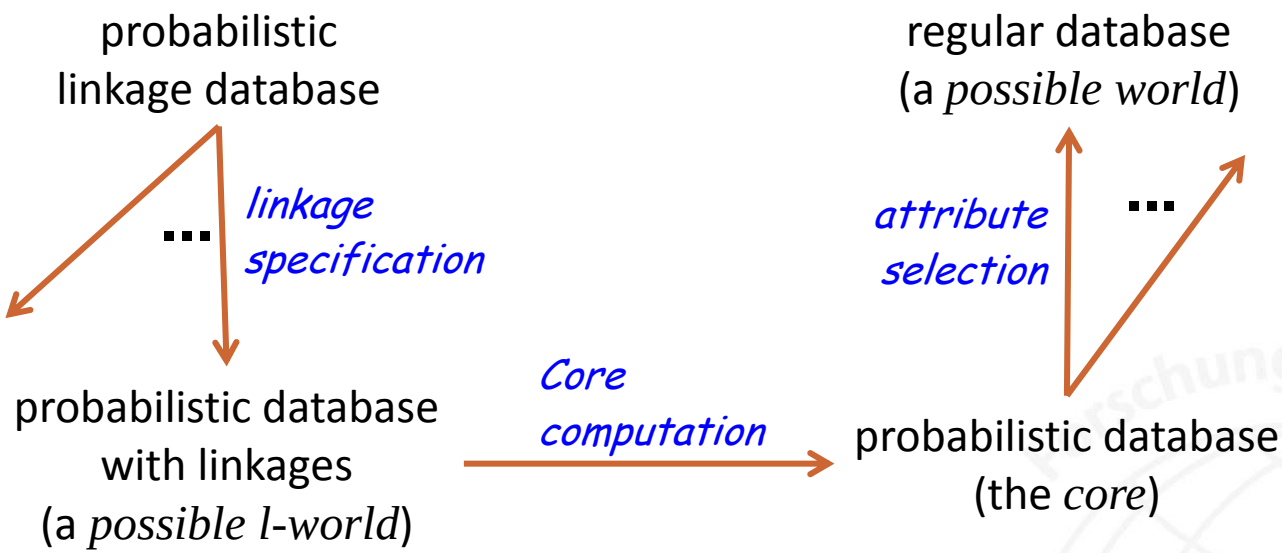
Some $\mathcal{L}^{sp}$ are invalid:
Example for $\rightarrow$ $\mathcal{L}=\{l_{e_1,e_2},\ l_{e_2,e_3},\ l_{e_1,e_3}\}$
$\mathcal{L}^{sp}=\{l_{e_1,e_2}, l_{e_1,e_3}\}$ is invalid — transitivity: $e_1 \equiv e_2 \equiv e_3$ AND $e_2 \neq e_3$

Possible l-world & Possible world

- Valid linkage specifications
  $\Rightarrow$ possible l-worlds
- Probabilities on linkages are eliminated
- BUT attribute probabilities are still present
- Generate the possible worlds
  (as performed in probabilistic databases)



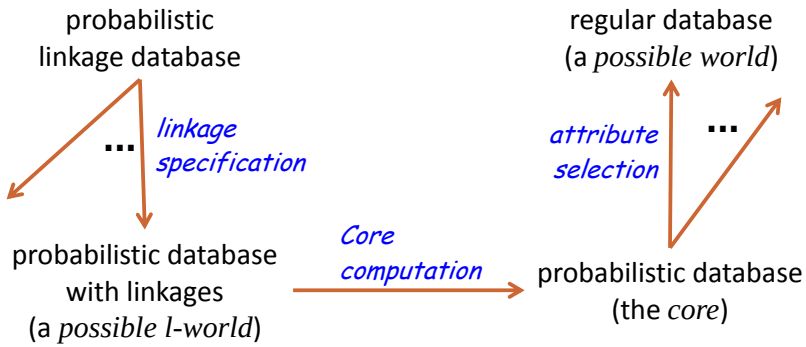
The different kinds of probabilistic databases



On-the-Fly Query Processing

Given a database and a query:

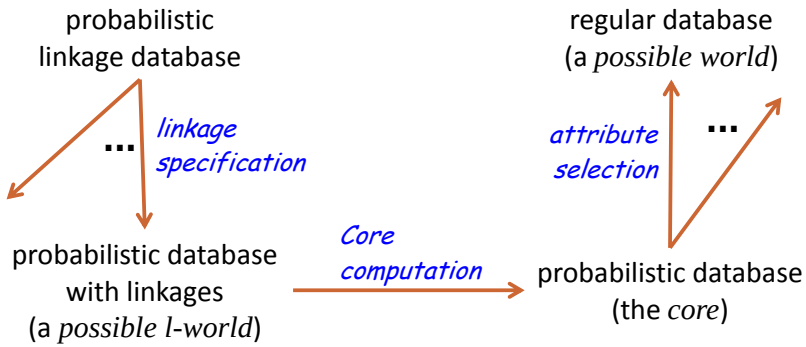
1. generate all possible *l*-worlds
2. identify and ignore invalid *l*-worlds
3. compute probability of each *l*-world
4. generate all possible worlds (for each *l*-world)
5. compute probability of each world
6. identify worlds satisfying query



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



↦ 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

[SI11] S. Staworko, E. Ioannou. *Management of inconsistencies in data integration*. Chapter to be included in Dagstuhl Follow-up Series on Data Exchange, Integration, and Streams, 2011.

[INNV10] E. Ioannou, W. Nejdl, C. Niederée, Y. Velegrakis. *On-the-Fly Entity-Aware Query Processing in the Presence of Linkage*. PVLDB, 3(1):429-438, 2010.

[Ioa09] E. Ioannou. *Entity-Aware Query Processing for Heterogeneous Data with Uncertainty and Correlations*. In Joint EDBT/ICDT Ph.D. Workshop, 2009.

Outline

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., $\{l_{e_1,e_2}, l_{e_1,e_3}, l_{e_4,e_5}\}$
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 *l*-worlds created as follows:

$$plw(\langle \mathcal{E}, \mathcal{L}, p^a, p^l \rangle) = \mathcal{L}^{sp}_{f_1} \times \mathcal{L}^{sp}_{f_2} \times \ldots \times \mathcal{L}^{sp}_{f_n}$$

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:

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

$\times$

Factor $f_2 = \{e_4,e_5\}$ for $\mathcal{L}_{f_2} = \{ l_{e_4,e_5} \}$	
$\mathcal{L}^{sp}_{f_2}(1) = \{ l_{e_4,e_5} \}$	0.8
$\mathcal{L}^{sp}_{f_2}(2) = \{\}$	$(1\text{-}0.8) = 0.2$

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:

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

$\times$

Factor $f_2 = \{e_4,e_5\}$ for $\mathcal{L}_{f_2} = \{ l_{e_4,e_5} \}$	
$\mathcal{L}^{sp}_{f_2}(1) = \{ l_{e_4,e_5} \}$	0.8
$\mathcal{L}^{sp}_{f_2}(2) = \{\}$	$(1\text{-}0.8) = 0.2$

Possible *I*-world	Required Merges	Probability
$I_1 = \{ l_{e_1,e_2},\ l_{e_1,e_3}, l_{e_4,e_5} \}$	$e_1 \equiv e_2 \equiv e_3,\ e_4 \equiv e_5$	$0.54 \times 0.8 = 0.432$
$I_2 = \{ l_{e_1,e_2},\ l_{e_1,e_3} \}$	$e_1 \equiv e_2 \equiv e_3,\ e_4,\ e_5$	$0.54 \times 0.2 = 0.108$
$I_3 = \{ l_{e_1,e_2},\ l_{e_4,e_5} \}$	$e_1 \equiv e_2,\ e_3,\ e_4 \equiv e_5$	$0.36 \times 0.8 = 0.288$
$I_4 = \{ l_{e_1,e_2} \}$	$e_1 \equiv e_2,\ e_3,\ e_4,\ e_5$	$0.36 \times 0.2 = 0.072$
$I_5 = \{ l_{e_1,e_3},\ l_{e_4,e_5} \}$	$e_1 \equiv e_3,\ e_2,\ e_4 \equiv e_5$	$0.06 \times 0.8 = 0.048$
$I_6 = \{ l_{e_1,e_3} \}$	$e_2,\ e_1 \equiv e_3,\ e_4,\ e_5$	$0.06 \times 0.2 = 0.012$
$I_7 = \{ l_{e_4,e_5} \}$	$e_1,\ e_2,\ e_3,\ e_4 \equiv e_5$	$0.04 \times 0.8 = 0.032$
$I_8 = \{\}$	$e_1,\ e_2,\ e_3,\ e_4,\ e_5$	$0.04 \times 0.2 = 0.008$

Deciding the Entity Merges

Exploit factors to avoid considering all the possible I-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 I-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"$\rangle$

Deciding the Entity Merges

Exploit factors to avoid considering all the possible I-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 I-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 $\langle f_1 - e_2, f_1 - e_2 \rangle$

Deciding the Entity Merges

Exploit factors to avoid considering all the possible I-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 $\langle f_1 - e_2, f_1 - e_2 \rangle$

$\mapsto$ merge(e_1, e_2), and merge(e_2)

Computing I-world probabilities

Probability given a query:

$$\prod_{i=1}^{m} Pr(\mathcal{L}_{f_i}^{sp} \mid c_m), \text{ where } c_m \text{ are the conditions describing a merge}$$

- 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

Possible worlds and their probabilities

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
- merge($e_1, \ldots, e_n$) = $\langle$ 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., $M = \{\{e_1.\alpha_i, e_1.\alpha_j, \ldots\}\}$
- merge($e_1, \ldots, e_n$) = $\langle$ id', $A \rangle$, where

$$A \subseteq (M_1 \times M_2 \times \ldots \times M_m) \cup \{ \alpha \mid \alpha \notin \cup_{i=1}^{m} M_i.\alpha \}$$

Possible worlds and their probabilities - Example

Consider exclusive attributes (name-value pair):

- starring: "Daniel Radcliffe"
- starring: "Emma Watson"

Figure shows the possible worlds for $merge(e_1, e_2)$

aid.	name	value	p
● a_{10}	starring	Daniel Radcliffe	0.7
◇ a_{11}	starring	Emma Watson	0.4
a_{12}	writer	J.K. Rowling	0.6
a_{13}	genre	Fantasy	0.6
● a_{20}	starring	Daniel Radcliffe	0.5
◇ a_{21}	starring	Emma Watson	0.9

Possible worlds and their probabilities - Example

Consider exclusive attributes (name-value pair):

- starring: "Daniel Radcliffe"
- starring: "Emma Watson"

Figure shows the possible worlds for $merge(e_1, e_2)$

	aid.	name	value	p
•	a_{10}	starring	Daniel Radcliffe	0.7
◇	a_{11}	starring	Emma Watson	0.4
	a_{12}	writer	J.K. Rowling	0.6
	a_{13}	genre	Fantasy	0.6
•	a_{20}	starring	Daniel Radcliffe	0.5
◇	a_{21}	starring	Emma Watson	0.9

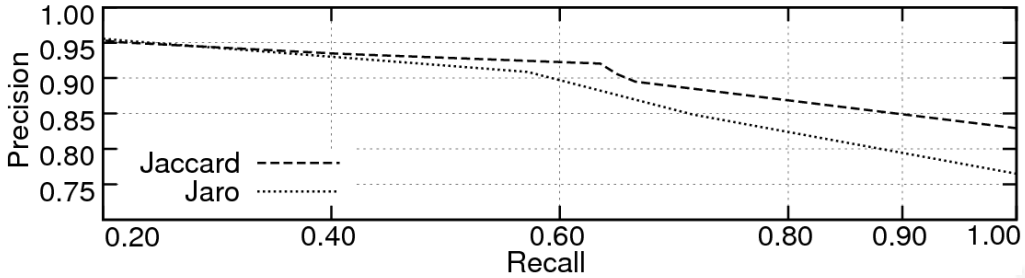
Possible Worlds

(1)	(2)	(3)	(4)
a_{10}	a_{20}	a_{10}	a_{20}
a_{11}	a_{11}	a_{21}	a_{21}
a_{12}	a_{12}	a_{12}	a_{12}
a_{13}	a_{13}	a_{13}	a_{13}

Experimental Evaluation - Influence of Linkages

Movie Dataset:

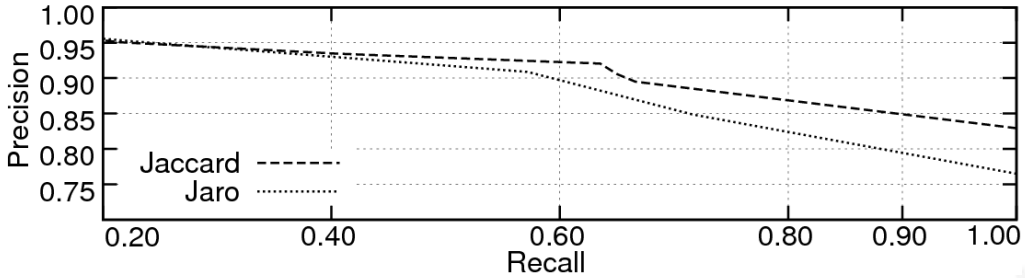
- 13,435 movies (23,182 IMDb, & 28,040 DBpedia)
- Two string similarity methods: Jaccard and Jaro



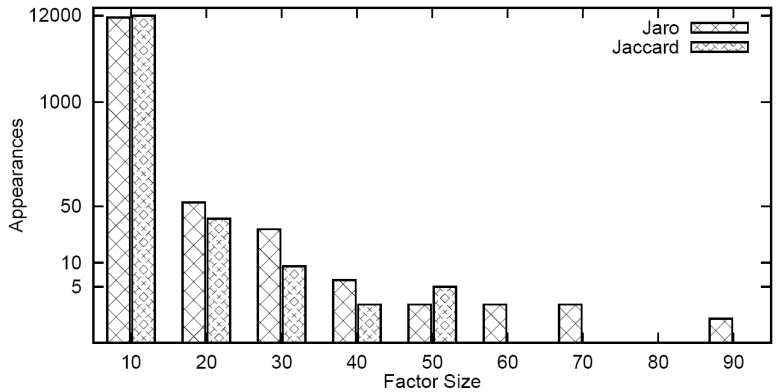
Experimental Evaluation - Influence of Linkages

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:

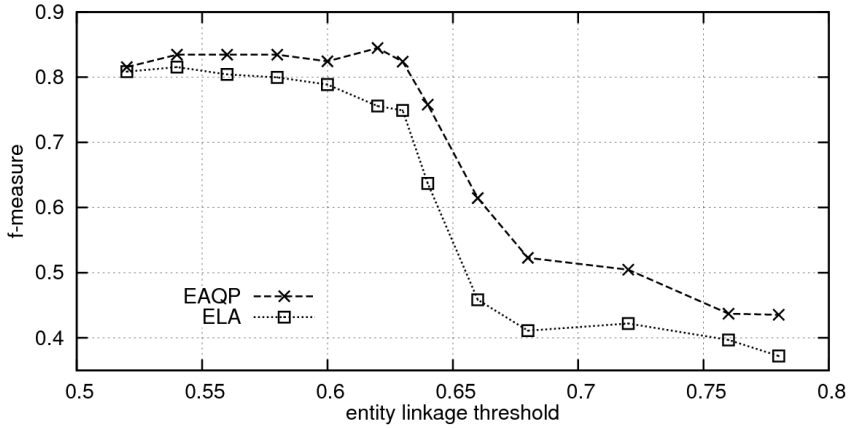
- Probabilistic entity linkages for publication authors
- 9,774 author descriptions that refer to 2,882 real world objects

Entity Linkages (under threshold t)					
t=0.52	t=0.58	t=0.62	t=0.68	t=0.72	t=0.78
12,440	12,012	10,775	6,394	5,985	4,184

Experimental Evaluation

Effectiveness:

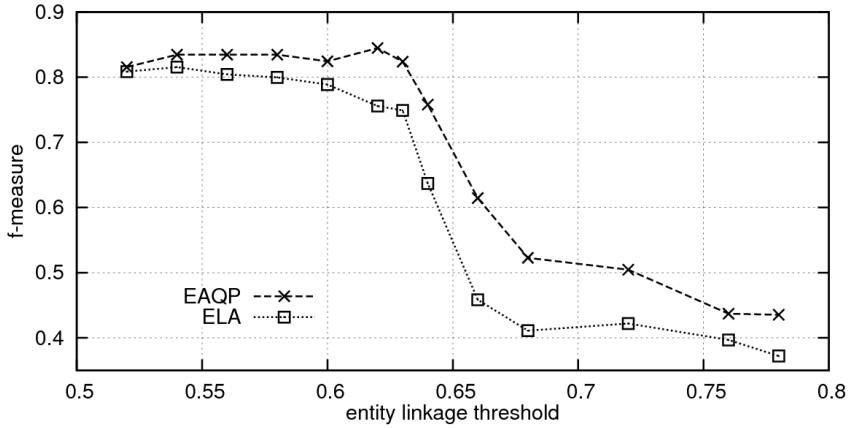
- F-measure: weighted harmonic mean of precision/recall
- EAQP exhibits a higher F-measure than ELA
- Higher difference for threshold values 0.65-0.75



Experimental Evaluation

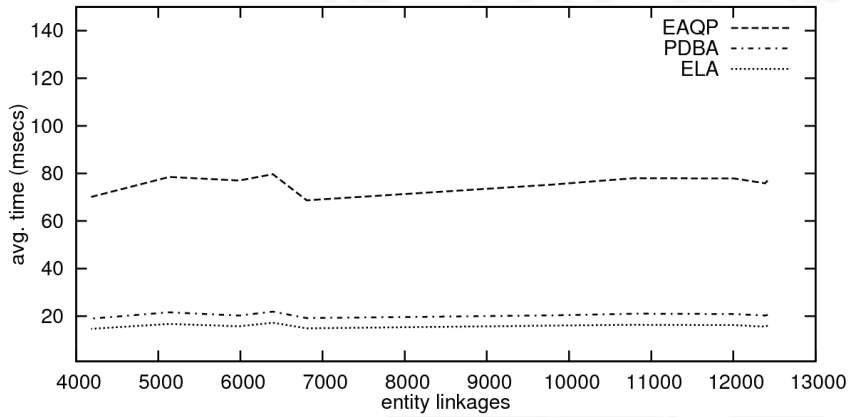
Effectiveness:

- 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

[INNV10] E. Ioannou, W. Nejdl, C. Niederée, Y. Velegrakis. *On-the-Fly Entity-Aware Query Processing in the Presence of Linkage*. PVLDB, 3(1):429-438, 2010.

[INNV11] E. Ioannou, W. Nejdl, C. Niederée, Y. Velegrakis. *LinkDB: A Probabilistic Linkage Database System*. In SIGMOD Conference (demo track), 2011.

Outline

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

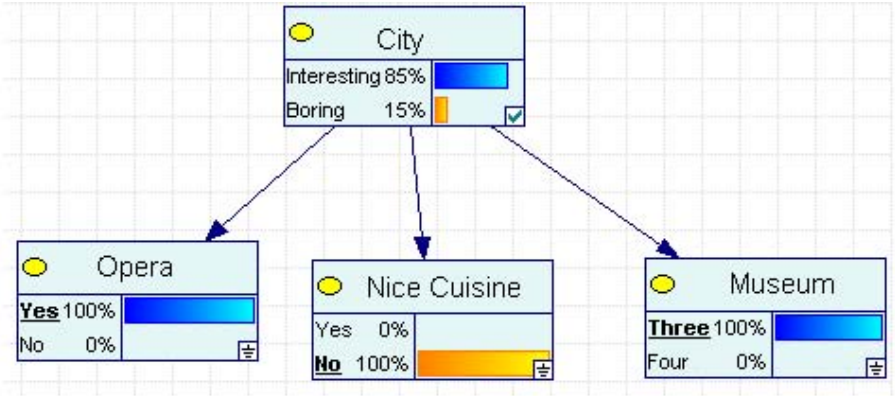
Nodes are accompanied with:

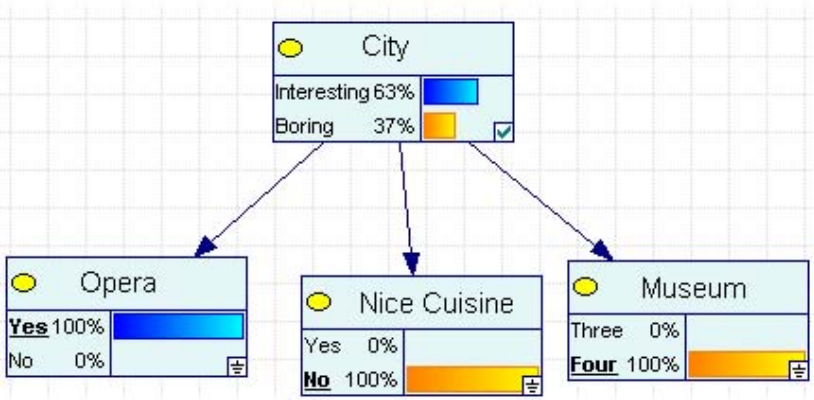
- Unconditional probability (no parents)
- Conditional probability (given parents)

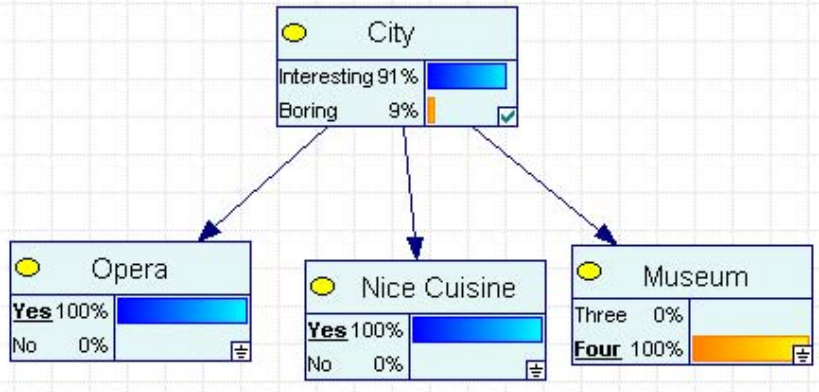
Probabilistic Inference:

Determines (given any new effects) the conditional probabilities of cause nodes

Bayesian Networks - Overview







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:

Effect Nodes:	(1) Evidence	(2) Direct-Rel.	(3) Deductive-Rel.
Cause nodes:			
(1) Linkage	√	√	
(2) Ded.-Rel.		√	√

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

  $P\,[\, e_{paper77.author1} = e_{paper127.author1}\,]$

  $\mapsto$dir-rel$(e_{paper77},\ e_{paper127})$

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:

$$\left.\begin{array}{l}\text{dir-rel}(e_{paper77}, e_{paper127}) \\ \text{dir-rel}(e_{paper77}, e_{email128})\end{array}\right\} \quad \text{ded-rel}(e_{paper127}, e_{email128})$$

- Add ded-rel node and cause-effect relationships
- Stop mechanism using evidence density

Incremental Computation of the Network

Step 1 - Add Evidence/Entity nodes

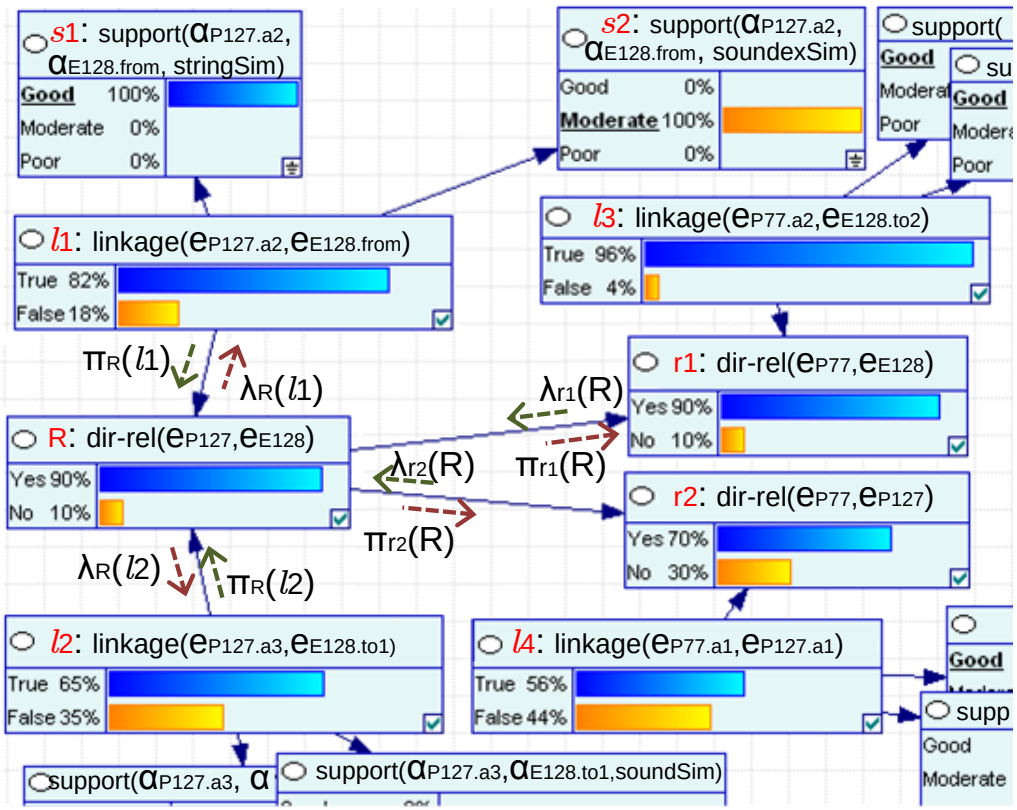
Step 2 - Add Direct-Relation nodes

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

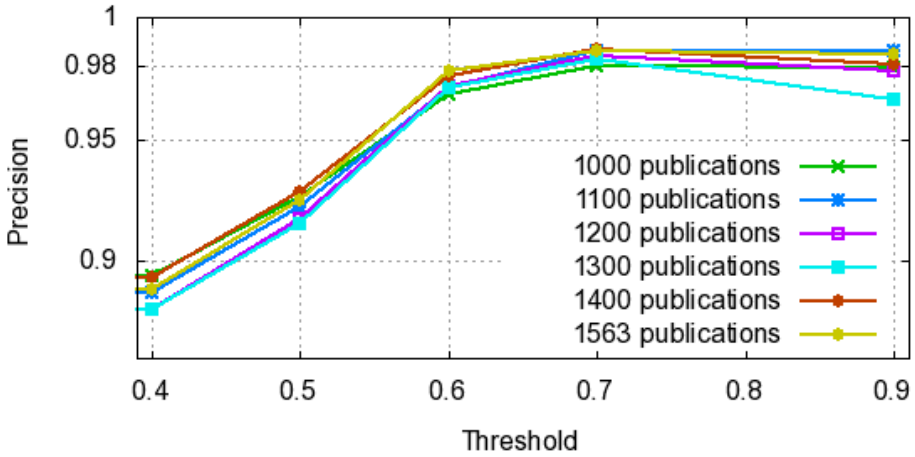
Name variants:

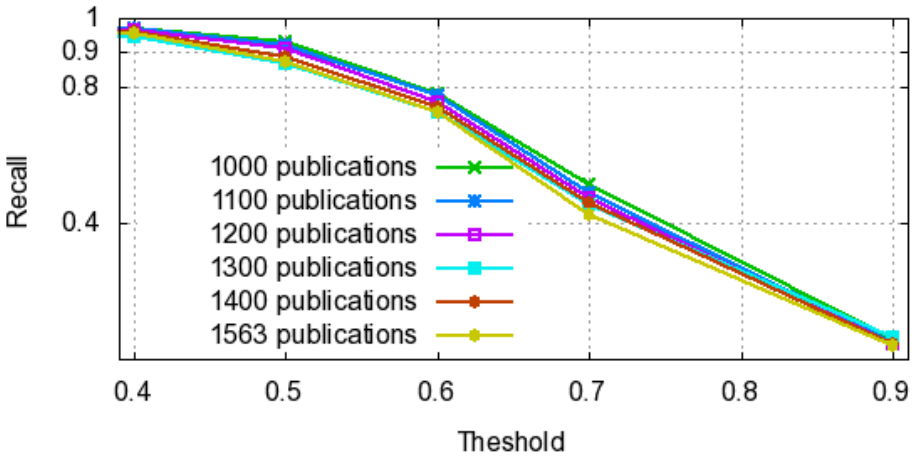
- Example $\mapsto$ "J. Antonisse" ; "Antonisse , H. J. " ; "Antonisse"
- Maximum is 88 different entities for the same object

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

[INN08] E. Ioannou, C. Niederée, W. Nejdl. *Probabilistic Entity Linkage for Heterogeneous Information Spaces*. In CAiSE, pages 556-570, 2008.

[IPSN10] E. Ioannou, O. Papapetrou, D. Skoutas, W. Nejdl. *Efficient Semantic-Aware Detection of Near Duplicate Resources*. In ESWC, pages 136-150, 2010.

[MPC+10] E. Minack, R. Paiu, S. Costache, G. Demartini, J. Gaugaz, E. Ioannou, P. Chirita, W. Nejdl. *Leveraging personal metadata for Desktop search: The Beagle++ system*. In Journal of Web Semantics, 8(1):37-54, 2010.

Outline

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

[ABS+06] Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha U. Nabar, Tomoe Sugihara, and Jennifer Widom.
Trio: A system for data, uncertainty, and lineage.
In *VLDB*, 2006.

[AFM06] Periklis Andritsos, Ariel Fuxman, and Renée J. Miller.
Clean answers over dirty databases: A probabilistic approach.
In *ICDE*, 2006.

[AKO07] Lyublena Antova, Christoph Koch, and Dan Olteanu.
10^{10^6} worlds and beyond: Efficient representation and processing of incomplete information.
In *ICDE*, 2007.

[CRF03] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg.
A comparison of string distance metrics for name-matching tasks.
In *IIWeb*, 2003.

[DH05] AnHai Doan and Alon Y. Halevy.
Semantic integration research in the database community: A brief survey.
AI Magazine, 2005.

[DHM05] Xin Dong, Alon Halevy, and Jayant Madhavan.
Reference reconciliation in complex information spaces.
In *SIGMOD*, 2005.

[DHY07] Xin Dong, Alon Y. Halevy, and Cong Yu.
Data integration with uncertainty.
In *VLDB*, 2007.

[DKP+09] Nilesh N. Dalvi, Ravi Kumar, Bo Pang, Raghu Ramakrishnan, Andrew Tomkins, Philip Bohannon, Sathiya Keerthi, and Srujana Merugu.
A web of concepts.
In *PODS*, 2009.

[DS04] Nilesh N. Dalvi and Dan Suciu.
Efficient query evaluation on probabilistic databases.

In *VLDB*, 2004.

[DS07] Nilesh N. Dalvi and Dan Suciu.
Management of probabilistic data: foundations and challenges.
In *PODS*, 2007.

[EIV07] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios.
Duplicate record detection: A survey.
IEEE Trans. Knowl. Data Eng., 2007.

[HFM06] Alon Y. Halevy, Michael J. Franklin, and David Maier.
Principles of dataspace systems.
In *PODS*, 2006.

[INN08] Ekaterini Ioannou, Claudia Niederée, and Wolfgang Nejdl.
Probabilistic entity linkage for heterogeneous information spaces.
In *CAiSE*, pages 556–570, 2008.

[INNV10] Ekaterini Ioannou, Wolfgang Nejdl, Claudia Niederée, and Yannis Velegrakis.
On-the-fly entity-aware query processing in the presence of linkage.
PVLDB, 3(1):429–438, 2010.

[INNV11] Ekaterini Ioannou, Wolfgang Nejdl, Claudia Niederée, and Yannis Velegrakis.
LinkDB: A probabilistic linkage database system.
In *SIGMOD Conference*, 2011.

[Ioa09] Ekaterini Ioannou.
Entity-aware query processing for heterogeneous data with uncertainty and correlations.
In *Joint EDBT/ICDT Ph.D. Workshop*, 2009.

[IPSN10] Ekaterini Ioannou, Odysseas Papapetrou, Dimitrios Skoutas, and Wolfgang Nejdl.
Efficient semantic-aware detection of near duplicate resources.
In *ESWC*, pages 136–150, 2010.

[KM06] Dmitri V. Kalashnikov and Sharad Mehrotra.
Domain-independent data cleaning via analysis of entity-relationship graph.
TODS, 2006.

[MPC+10] Enrico Minack, Raluca Paiu, Stefania Costache, Gianluca Demartini, Julien Gaugaz, Ekaterini Ioannou, Paul-Alexandru Chirita, and Wolfgang Nejdl.
Leveraging personal metadata for desktop search: The beagle++ system.
Journal of Web Semantics, 8(1):37–54, 2010.

[OS99] Aris M. Ouksel and Amit P. Sheth.
Semantic interoperability in global information systems: A brief introduction to the research area and the special section.
SIGMOD, 1999.

[RDS07] Christopher Re, Nilesh N. Dalvi, and Dan Suciu.
Efficient top-k query evaluation on probabilistic data.
In *ICDE*, 2007.

[RS08] Christopher Re and Dan Suciu.
Managing probabilistic data with mystiq: The can-do, the could-do, and the can't-do.
In *SUM*, 2008.

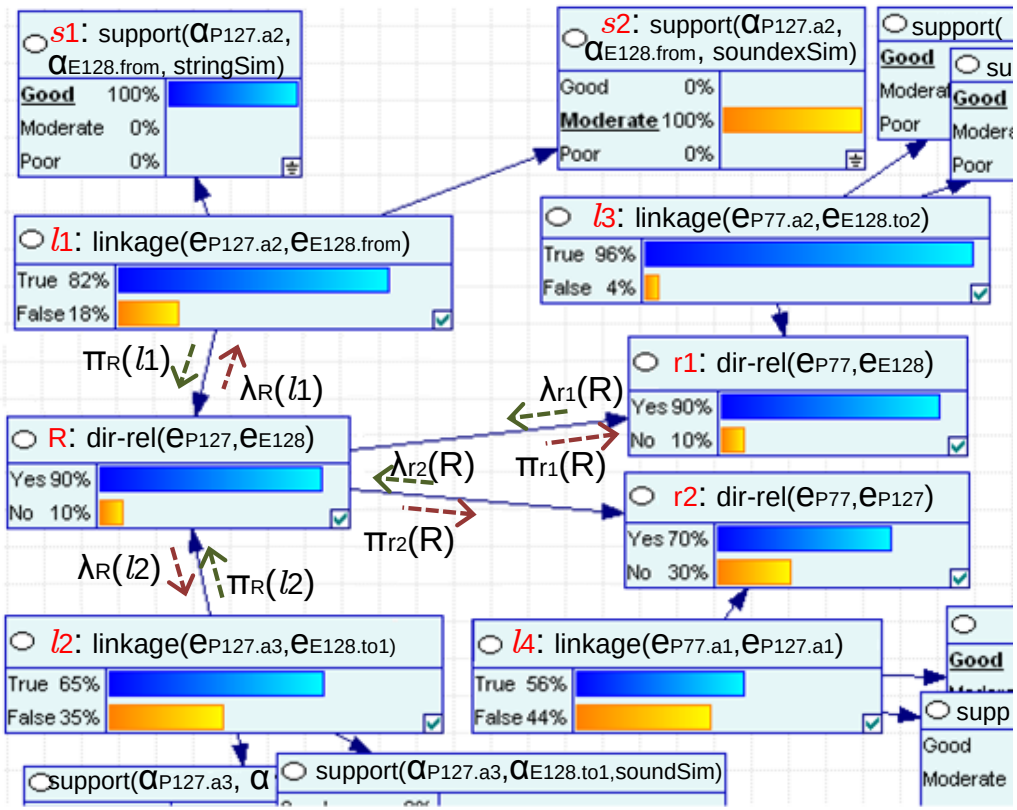
[SD07] Prithviraj Sen and Amol Deshpande.
Representing and querying correlated tuples in probabilistic databases.
In *ICDE*, 2007.

[SI11] Slawek Staworko and Ekaterini Ioannou.
Management of inconsistencies in data integration.
Chapter to be included in Dagstuhl Follow-up Series on Data Exchange, Integration, and Streams, 2011.

[Vel08] Yannis Velegrakis.
On the importance of updates in information integration and data exchange systems.
In *DBISP2P*, 2008.

[WMK+09] Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina.
Entity resolution with iterative blocking.
In *SIGMOD Conference*, 2009.

Example



When R is activated:

Receives $\lambda_{r_1}(R)$, $\lambda_{r_2}(R)$, $\pi_R(I1)$, $\pi_R(I2)$

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

$$\lambda(R) = \lambda_{r_1}(R)\lambda_{r_2}(R)$$
$$\pi(R) = P(R|I1, I2)\pi_R(I1)\pi_R(I2)$$

Sends message to parent nodes:

$$\lambda_R(I1) = P(R|I1, I2)\pi_R(I2)\lambda(R)$$

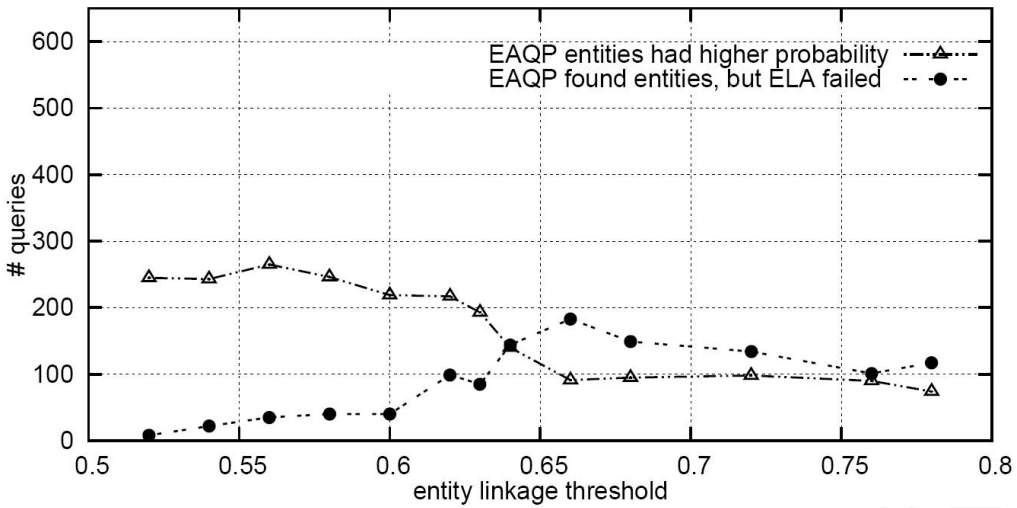
Sends messages to children nodes:

$$\pi_{r2}(R) = \pi(R)\lambda_{r1}(R)$$

Experimental Evaluation

Improvements over ELA:

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



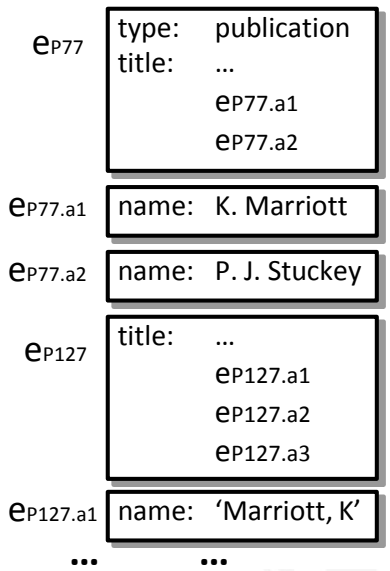
Example

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

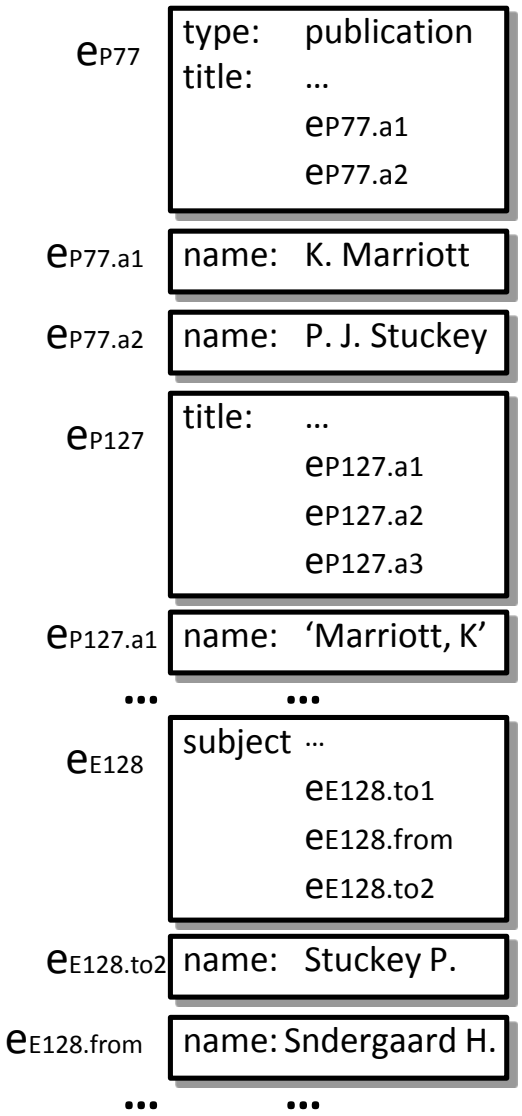
metadata for publ. #77 (M(r77))={...,
<file:///P77, type, publication>,
<file:///P77, title, ... >,
<file:///P77/a1, name, K. Marriott>,
<file:///P77/a2, name, P. J. Stuckey>}

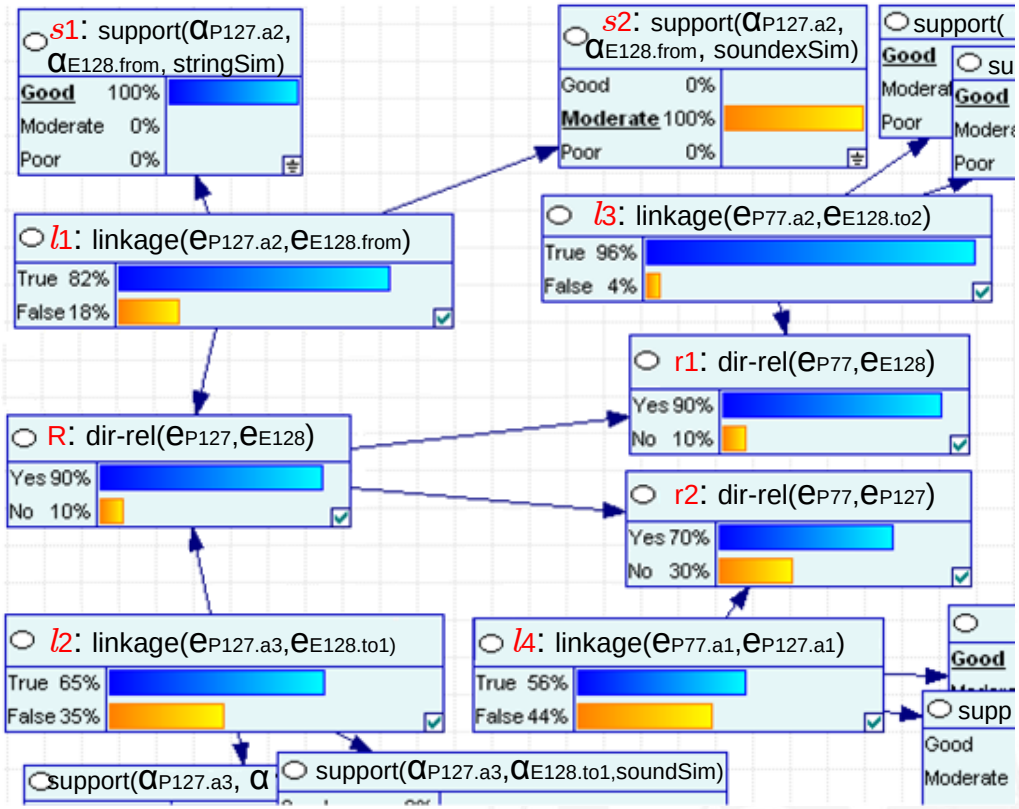
metadata for publ. #127 (M(r127))={...,
<file:///P127/a1, name, `Marriott, K'>,
<file:///P127/a2, name, `Sndergaard, H' >,
<file:///P127/a3, name, `Kelly, A'>}

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


eP77
type: publication
title: ...
eP77.a1
eP77.a2
eP77.a1
name: K. Marriott
eP77.a2
name: P. J. Stuckey
eP127
title: ...
eP127.a1
eP127.a2
eP127.a3
eP127.a1
name: 'Marriott, K'
eE128
subject ...
eE128.to1
eE128.from
eE128.to2
eE128.to2
name: Stuckey P.
eE128.from
name: Sndergaard H.


s1: support(αP127.a2, αE128.from, stringSim)
Good 100%
Moderate 0%
Poor 0%
s2: support(αP127.a2, αE128.from, soundexSim)
Good 0%
Moderate 100%
Poor 0%
support(
Good
Poor
l1: linkage(eP127.a2,eE128.from)
True 82%
False 18%
l3: linkage(eP77.a2,eE128.to2)
True 96%
False 4%
r1: dir-rel(eP77,eE128)
Yes 90%
No 10%
R: dir-rel(eP127,eE128)
Yes 90%
No 10%
r2: dir-rel(eP77,eP127)
Yes 70%
No 30%
l2: linkage(eP127.a3,eE128.to1)
True 65%
False 35%
l4: linkage(eP77.a1,eP127.a1)
True 56%
False 44%
Good
supp
Good
Moderate
support(αP127.a3, α
support(αP127.a3,αE128.to1,soundSim)