Probabilistic Reasoning

Thomas Lukasiewicz

Department of Computer Science, University of Oxford, UK
thomas.lukasiewicz@cs.ox.ac.uk
Outline

1. Probabilistic Description Logics
   - Motivation
   - Probabilistic Logics
   - P-$SHI\bar{F}(D)$ and P-$SHOIN(D)$

2. Probabilistic Datalog+/-
   - Datalog+/-
   - Markov Logic Networks
   - Probabilistic Datalog+/-

3. Probabilistic Ontological Data Exchange
   - Motivation and Overview
   - (Probabilistic) Ontological Data Exchange
   - Complexity Results
Outline

1. Probabilistic Description Logics
   - Motivation
   - Probabilistic Logics
   - \( P-\text{SHIF}(D) \) and \( P-\text{SHOIN}(D) \)

2. Probabilistic Datalog+/–
   - Datalog+/–
   - Markov Logic Networks
   - Probabilistic Datalog+/–

3. Probabilistic Ontological Data Exchange
   - Motivation and Overview
   - (Probabilistic) Ontological Data Exchange
   - Complexity Results
Probabilistic Ontologies

Generalization of classical ontologies by probabilistic knowledge.

Main types of encoded probabilistic knowledge:

- Terminological probabilistic knowledge about concepts and roles:
  “Birds fly with a probability of at least 0.95”.

- Assertional probabilistic knowledge about instances of concepts and roles:
  “Tweety is a bird with a probability of at least 0.9”.
Use of Probabilistic Ontologies

- In medicine, biology, defense, astronomy, ...
- In the Semantic Web:
  - Quantifying the degrees of overlap between concepts, to use them in Semantic Web applications: information retrieval, personalization, recommender systems, ...
  - Information retrieval, for an increased recall (e.g., Udrea et al.: Probabilistic ontologies and relational databases. In *Proc. CoopIS/DOA/ODBASE-2005*).
  - Ontology matching (e.g., Mitra et al.: OMEN: A probabilistic ontology mapping tool. In *Proc. ISWC-2005*).
  - Probabilistic data integration, especially for handling ambiguous and inconsistent pieces of information.
Description Logics: Key Ideas

Description logics model a domain of interest in terms of concepts and roles, which represent classes of individuals and binary relations between classes of individuals, respectively.

A description logic knowledge base encodes in particular subset relationships between concepts, subset relationships between roles, the membership of individuals to concepts, and the membership of pairs of individuals to roles.

Here, description logic knowledge bases in $\mathcal{SHIF}(D)$ and $\mathcal{SHOIN}(D)$ (which are the DLs behind OWL Lite and OWL DL, respectively).
Example

Description logic knowledge base $L$ for an online store:

(1) $\text{Textbook} \sqsubseteq \text{Book}$; (2) $\text{PC} \sqcup \text{Laptop} \sqsubseteq \text{Electronics}$; $\text{PC} \sqsubseteq \neg \text{Laptop}$;
(3) $\text{Book} \sqcup \text{Electronics} \sqsubseteq \text{Product}$; $\text{Book} \sqsubseteq \neg \text{Electronics}$;
(4) $\text{Sale} \sqsubseteq \text{Product}$;
(5) $\text{Product} \sqsubseteq \geq 1 \text{ related}$; (6) $\geq 1 \text{ related} \sqcup \geq 1 \text{ related}^- \sqsubseteq \text{Product}$;
(7) $\text{related} \sqsubseteq \text{related}^-; \text{related}^- \sqsubseteq \text{related}$;
(8) $\text{Textbook}(\text{tb}_\text{ai})$; $\text{Textbook}(\text{tb}_\text{lp})$; (9) $\text{related}(\text{tb}_\text{ai}, \text{tb}_\text{lp})$;
(10) $\text{PC}(\text{pc}_\text{ibm})$; $\text{PC}(\text{pc}_\text{hp})$; (11) $\text{related}(\text{pc}_\text{ibm}, \text{pc}_\text{hp})$;
(12) $\text{provides}(\text{ibm}, \text{pc}_\text{ibm})$; $\text{provides}(\text{hp}, \text{pc}_\text{hp})$. 
Probabilistic Logics: Key Ideas

- Integration of (propositional) logic- and probability-based representation and reasoning formalisms.

- Reasoning from logical constraints and interval restrictions for conditional probabilities (also called *conditional constraints*).

- Reasoning from convex sets of probability distributions.

- Model-theoretic notion of logical entailment.
Syntax of Probabilistic Knowledge Bases

- Finite nonempty set of basic events $\Phi = \{p_1, \ldots, p_n\}$.
- Event $\phi$: Boolean combination of basic events
- Logical constraint $\psi \leftarrow \phi$: events $\psi$ and $\phi$: “$\phi$ implies $\psi$”.
- Conditional constraint $(\psi|\phi)[l, u]$: events $\psi$ and $\phi$, and $l, u \in [0, 1]$: “conditional probability of $\psi$ given $\phi$ is in $[l, u]$”.
- Probabilistic knowledge base $KB = (L, P)$:
  - finite set of logical constraints $L$,
  - finite set of conditional constraints $P$. 
Probabilistic knowledge base $KB = (L, P)$:

- $L = \{bird \leftarrow eagle\}$:
  
  “All eagles are birds”.

- $P = \{(have\_legs \mid bird)[1, 1], (fly \mid bird)[0.95, 1]\}$:

  “All birds have legs”.

  “Birds fly with a probability of at least 0.95”.
Semantics of Probabilistic Knowledge Bases

- **World $I$:** truth assignment to all basic events in $\Phi$.
- $\mathcal{I}_\Phi$: all worlds for $\Phi$.
- **Probabilistic interpretation $\Pr$:** probability function on $\mathcal{I}_\Phi$.
- $\Pr(\phi)$: sum of all $\Pr(I)$ such that $I \in \mathcal{I}_\Phi$ and $I \models \phi$.
- $\Pr(\psi \mid \phi)$: if $\Pr(\phi) > 0$, then $\Pr(\psi \mid \phi) = \Pr(\psi \land \phi) / \Pr(\phi)$.
- **Truth under $\Pr$:**
  - $\Pr \models \psi \iff \Pr(\psi \land \phi) = \Pr(\phi) $ (iff $\Pr(\psi \iff \phi) = 1$).
  - $\Pr \models (\psi \mid \phi)[l, u] \iff \Pr(\psi \land \phi) \in [l, u] \cdot \Pr(\phi)$ (iff either $\Pr(\phi) = 0$ or $\Pr(\psi \mid \phi) \in [l, u]$).
Example

- Set of basic propositions $\Phi = \{\text{bird, fly}\}$.
- $\mathcal{I}_\Phi$ contains exactly the worlds $I_1, I_2, I_3,$ and $I_4$ over $\Phi$:

<table>
<thead>
<tr>
<th></th>
<th>fly</th>
<th>¬fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>bird</td>
<td>$I_1$</td>
<td>$I_2$</td>
</tr>
<tr>
<td>¬bird</td>
<td>$I_3$</td>
<td>$I_4$</td>
</tr>
</tbody>
</table>

- Some probabilistic interpretations:

<table>
<thead>
<tr>
<th></th>
<th>fly</th>
<th>¬fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr$_1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bird</td>
<td>19/40</td>
<td>1/40</td>
</tr>
<tr>
<td>¬bird</td>
<td>10/40</td>
<td>10/40</td>
</tr>
<tr>
<td>Pr$_2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bird</td>
<td>0</td>
<td>1/3</td>
</tr>
<tr>
<td>¬bird</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

- $\Pr_1(\text{fly} \land \text{bird}) = 19/40$ and $\Pr_1(\text{bird}) = 20/40$.
- $\Pr_2(\text{fly} \land \text{bird}) = 0$ and $\Pr_2(\text{bird}) = 1/3$.
- $\neg\text{fly} \Leftarrow \text{bird}$ is false in $\Pr_1$, but true in $\Pr_2$.
- $(\text{fly} \mid \text{bird})[.95, 1]$ is true in $\Pr_1$, but false in $\Pr_2$. 
Satisfiability and Logical Entailment

- Pr is a model of $KB = (L, P)$ iff $Pr \models F$ for all $F \in L \cup P$.
- $KB$ is satisfiable iff a model of $KB$ exists.
- $KB \models (\psi|\phi)[l, u]$: $(\psi|\phi)[l, u]$ is a logical consequence of $KB$ iff every model of $KB$ is also a model of $(\psi|\phi)[l, u]$.
- $KB \models_{\text{tight}} (\psi|\phi)[l, u]$: $(\psi|\phi)[l, u]$ is a tight logical consequence of $KB$ iff $l$ (resp., $u$) is the infimum (resp., supremum) of $Pr(\psi|\phi)$ subject to all models $Pr$ of $KB$ with $Pr(\phi) > 0$. 
Example

- Probabilistic knowledge base:
  \[ KB = (\{bird \equiv eagle\}, \{(have\_legs \mid bird)[1, 1], (fly \mid bird)[0.95, 1]\}) \].

- \( KB \) is satisfiable, since
  \[ \Pr \text{ with } \Pr(bird \land eagle \land have\_legs \land fly) = 1 \text{ is a model.} \]

- Some conclusions under logical entailment:
  \[ KB \models (have\_legs \mid bird)[0.3, 1], \quad KB \models (fly \mid bird)[0.6, 1]. \]

- Tight conclusions under logical entailment:
  \[ KB \models_{\text{tight}} (have\_legs \mid bird)[1, 1], \quad KB \models_{\text{tight}} (fly \mid bird)[0.95, 1], \quad KB \models_{\text{tight}} (have\_legs \mid eagle)[1, 1], \quad KB \models_{\text{tight}} (fly \mid eagle)[0, 1]. \]
Towards Stronger Notions of Entailment

Problem: Inferential weakness of logical entailment.

Solutions:

- **Probability selection techniques:** Perform inference from a representative distribution of the encoded convex set of distributions rather than the whole set, e.g.,
  - distribution of maximum entropy,
  - distribution in the center of mass.

- **Probabilistic default reasoning:** Perform constraining rather than conditioning and apply techniques from default reasoning to resolve local inconsistencies.

- **Probabilistic independencies:** Further constrain the convex set of distributions by probabilistic independencies. (⇒ adds nonlinear equations to linear constraints)
Logical vs. Lexicographic Entailment

Probabilistic knowledge base:

\[ KB = (\{ bird \leftarrow eagle \}, \{ (have\_legs \mid bird)[1, 1], (fly \mid bird)[0.95, 1] \}). \]

Tight conclusions under logical entailment:

\[ KB \models_{\text{tight}} (have\_legs \mid bird)[1, 1], \quad KB \models_{\text{tight}} (fly \mid bird)[0.95, 1], \]
\[ KB \models_{\text{tight}} (have\_legs \mid eagle)[1, 1], \quad KB \models_{\text{tight}} (fly \mid eagle)[0, 1]. \]

Tight conclusions under probabilistic lexicographic entailment:

\[ KB \models_{\text{t} \text{ lex}} (have\_legs \mid bird)[1, 1], \quad KB \models_{\text{t} \text{ lex}} (fly \mid bird)[0.95, 1], \]
\[ KB \models_{\text{t} \text{ lex}} (have\_legs \mid eagle)[1, 1], \quad KB \models_{\text{t} \text{ lex}} (fly \mid eagle)[0.95, 1]. \]
Probabilistic knowledge base:

\[ KB = (\{ bird \leftarrow penguin \}, \{(have\_legs \mid bird)[1, 1], \\
(fly \mid bird)[1, 1], (fly \mid penguin)[0, 0.05]\}). \]

Tight conclusions under logical entailment:

\[ KB \models_{tight} (have\_legs \mid bird)[1, 1], KB \models_{tight} (fly \mid bird)[1, 1], \\
KB \models_{tight} (have\_legs \mid penguin)[1, 0], KB \models_{tight} (fly \mid penguin)[1, 0]. \]

Tight conclusions under probabilistic lexicographic entailment:

\[ KB \models_{tight}^{lex} (have\_legs \mid bird)[1, 1], KB \models_{tight}^{lex} (fly \mid bird)[1, 1], \\
KB \models_{tight}^{lex} (have\_legs \mid penguin)[1, 1], KB \models_{tight}^{lex} (fly \mid penguin)[0, 0.05]. \]
Probabilistic knowledge base:

\[
KB = (\{bird \iff penguin\}, \{(have\_legs \mid bird)[0.99, 1],
(fly \mid bird)[0.95, 1], (fly \mid penguin)[0, 0.05]\}).
\]

Tight conclusions under logical entailment:

\[
KB \models_{t\text{ight}} (have\_legs \mid bird)[0.99, 1], \ KB \models_{t\text{ight}} (fly \mid bird)[0.95, 1],
KB \models_{t\text{ight}} (have\_legs \mid penguin)[0, 1], \ KB \models_{t\text{ight}} (fly \mid penguin)[0, 0.05].
\]

Tight conclusions under probabilistic lexicographic entailment:

\[
KB \models_{t\text{ight}}^{\text{lex}} (have\_legs \mid bird)[0.99, 1], \ KB \models_{t\text{ight}}^{\text{lex}} (fly \mid bird)[0.95, 1],
KB \models_{t\text{ight}}^{\text{lex}} (have\_legs \mid penguin)[0.99, 1], \ KB \models_{t\text{ight}}^{\text{lex}} (fly \mid penguin)[0, 0.05].
\]
P-$SHIF(D)$ and P-$SHOIN(D)$: Key Ideas

- probabilistic generalization of the description logics $SHIF(D)$ and $SHOIN(D)$ behind OWL Lite and OWL DL, respectively
- terminological probabilistic knowledge about concepts and roles
- assertional probabilistic knowledge about instances of concepts and roles
- terminological probabilistic inference based on lexicographic entailment in probabilistic logic (stronger than logical entailment)
- assertional probabilistic inference based on lexicographic entailment in probabilistic logic (for combining assertional and terminological probabilistic knowledge)
- terminological and assertional probabilistic inference problems reduced to sequences of linear optimization problems
Example

Standard terminological knowledge:

(1) \textit{MalePacemakerPatient} \sqsubseteq \textit{PacemakerPatient},
    \textit{FemalePacemakerPatient} \sqsubseteq \textit{PacemakerPatient},
(2) \textit{MalePacemakerPatient} \sqsubseteq \neg \textit{FemalePacemakerPatient},
(3) \textit{PacemakerPatient} \sqsubseteq \textit{HeartPatient},
(4) \exists \textit{HasIllnessSymptom}. \top \sqsubseteq \textit{HeartPatient},
    \exists \textit{HasIllnessSymptom}^\neg \cdot \top \sqsubseteq \textit{IllnessSymptom},
(5) \textit{HeartPatient}(\textit{Tom}),
(6) \textit{MalePacemakerPatient}(\textit{John}),
(7) \textit{FemalePacemakerPatient}(\textit{Maria}),
(8) \textit{HasIllnessSymptom}(\textit{John}, \textit{Arrhythmia}),
    \textit{HasIllnessSymptom}(\textit{John}, \textit{ChestPain}),
    \textit{HasIllnessSymptom}(\textit{John}, \textit{BreathingDifficulties}),
    \textit{HasIllnessStatus}(\textit{John}, \textit{Advanced}).
Example

Terminological default and probabilistic knowledge:

(9) \( (\text{HighBloodPressure} \mid \text{HeartPatient})[1, 1], \)
(10) \( (\neg \text{HighBloodPressure} \mid \text{PacemakerPatient})[1, 1], \)
(11) \( (\text{MalePacemakerPatient} \mid \text{PacemakerPatient})[0.4, 1], \)
(12) \( (\exists \text{HasHealthInsurance. PrivateHealthInsurance} \mid \text{HeartPatient})[0.9, 1], \)
(13) \( (\exists \text{HasIllnessSymptom. \{Arrhythmia\}} \mid \text{PacemakerPatient})[0.98, 1], \)
(\( \exists \text{HasIllnessSymptom. \{ChestPain\}} \mid \text{PacemakerPatient})[0.9, 1], \)
(\( \exists \text{HasIllnessSymptom. \{BreathingDifficulties\}} \mid \text{PacemakerPatient})[0.6, 1]. \)
Example

Assertional probabilistic knowledge:

For individual Tom:

\[(14) \ (\text{PacemakerPatient} \mid \top) [0.8, 1].\]

For individual Maria:

\[(15) \ (∃ \text{HasIllnessSymptom.}\{\text{BreathingDifficulties}\} \mid \top) [0.6, 1],\]
\[(16) \ (∃ \text{HasIllnessSymptom.}\{\text{ChestPain}\} \mid \top) [0.9, 1],\]
\[(17) \ (∃ \text{HasIllnessStatus.}\{\text{Final}\} \mid \top) [0.2, 0.8].\]
## Complexity Results

<table>
<thead>
<tr>
<th></th>
<th>$P-DL$-Lite</th>
<th>$P-SHIF(D)$</th>
<th>$P-SHOIN(D)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SAT}$</td>
<td>NP</td>
<td>EXP</td>
<td>NEXP</td>
</tr>
<tr>
<td>$\text{PTCON}$</td>
<td>NP</td>
<td>EXP</td>
<td>NEXP</td>
</tr>
<tr>
<td>$\text{PKBCON}$</td>
<td>NP</td>
<td>EXP</td>
<td>NEXP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$P-DL$-Lite</th>
<th>$P-SHIF(D)$</th>
<th>$P-SHOIN(D)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{TLOGEnt}$</td>
<td>$\text{FP}^{NP}$</td>
<td>$\text{FEXP}$</td>
<td>in $\text{FP}^{NEXP}$</td>
</tr>
<tr>
<td>$\text{TLEXEnt}$</td>
<td>$\text{FP}^{NP}$</td>
<td>$\text{FEXP}$</td>
<td>in $\text{FP}^{NEXP}$</td>
</tr>
</tbody>
</table>
References

Outline

1. Probabilistic Description Logics
   - Motivation
   - Probabilistic Logics
   - $P$-SHIF($D$) and $P$-SHOIN($D$)

2. Probabilistic Datalog+/-
   - Datalog+/-
   - Markov Logic Networks
   - Probabilistic Datalog+/-

3. Probabilistic Ontological Data Exchange
   - Motivation and Overview
   - (Probabilistic) Ontological Data Exchange
   - Complexity Results
Probabilistic Datalog+/–: Key Ideas

- Probabilistic Datalog+/– ontologies combine “classical” Datalog+/– with Markov logic networks (MLNs).

- The basic idea is that formulas (TGDs, EGDs, and NCs) are annotated with a set of probabilistic events.

- Event annotations mean that the formula in question only applies when the associated event holds.

- The probability distribution associated with the events is described in the MLN.

- Key computational problems: answering ranking queries, conjunctive queries, and threshold queries.

- Application in data extraction from the Web, where Datalog+/– is used as data extraction language (DIADEM).
Consider the problem of *entity extraction* over the following text snippet:

Fifty Shades novels drop in sales after 22 weeks, according to trade magazine The Bookseller. According to the Bookseller, £29.3m was spent at UK booksellers between 15 and 22 September - a rise of £700,000 on the previous week.
Datalog+/−: Encoding Ontologies in Datalog

Plain Datalog allows for encoding some ontological axioms:

- concept inclusion axioms:
  \[ \text{person}(X) \leftarrow \text{employee}(X) \ \text{iff} \ \text{employee} \sqsubseteq \text{person}; \]

- role inclusion axioms:
  \[ \text{manages}(X, Y) \leftarrow \text{reportsTo}(Y, X) \ \text{iff} \ \text{reportsTo}^{-1} \sqsubseteq \text{manages}; \]

- concept and role membership axioms:
  \[ \text{person}(\text{John}) \leftarrow \ \text{iff} \ \text{person}(\text{John}); \]
  \[ \text{manages}(\text{Bill}, \text{John}) \leftarrow \ \text{iff} \ \text{manages}(\text{Bill}, \text{John}). \]

- transitivity axioms:
  \[ \text{manages}(X, Y) \leftarrow \text{manages}(X, Z), \text{manages}(Z, Y) \ \text{iff} \ \text{(Trans manages)} \]
However, it cannot express other important ontological axioms:

- concept inclusion axioms involving existential restrictions on roles in the head:
  
  $\text{Scientist} \sqsubseteq \exists \text{isAuthorOf}$;

- concept inclusion axioms stating concept disjointness:
  
  $\text{JournalPaper} \sqsubseteq \lnot \text{ConferencePaper}$;

- functionality axioms:
  
  (funct $\text{hasFirstAuthor}$).

Question: Can Datalog be extended in such a way that it can be used as ontology language?
Answer: Yes, by introducing:

- **tuple-generating dependencies (TGDs):**
  \[ \forall X \forall Y \exists Z \; \psi(X, Z) \leftarrow \Phi(X, Y), \]
  where \( \Phi(X, Y) \) and \( \psi(X, Z) \) are conjunctions of atoms;

  **Example:** \( \exists P \; \text{directs}(M, P) \leftarrow \text{manager}(M); \)

- **negative constraints:**
  \[ \forall X \bot \leftarrow \Phi(X), \]
  where \( \Phi(X) \) is a conjunction of atoms;

  **Example:** \( \bot \leftarrow c(X), c'(X); \)

- **equality-generating dependencies (EGDs):**
  \[ \forall X \; X_i = X_j \leftarrow \Phi(X), \]
  where \( X_i, X_j \in X \), and \( \Phi(X) \) is a conjunction of atoms

  **Example:** \( Y = Z \leftarrow r_1(X, Y), r_2(Y, Z). \)
The Chase

Given:

- \( D \): database over \( \text{dom}(D) \).
- \( \Sigma \): set of TGDs and/or EGDs

Question: How do we perform query answering?

Answer: Via the chase: If \( D \not\models \Sigma \), then

- either \( D \cup \Sigma \) is unsatisfiable due to a “hard” EGD violation, or
- the rules in \( \Sigma \) can be enforced via the chase by
  - adding facts in order to satisfy TGDs, where null values are introduced for \( \exists \)-variables
  - equating nulls with other nulls or with \( \text{dom}(D) \) elements in order to satisfy EGDs.
The Chase is a Universal Model

For each other model $M$ of $D$ and $\Sigma$, there is a homomorphism from chase($D, \Sigma$) to $M$.

⇒ conjunctive queries to $D \cup \Sigma$ can be evaluated on chase($D, \Sigma$):

$$D \cup \Sigma \models Q \text{ iff } \text{chase}(D, \Sigma) \models Q$$
Facts about the Chase

- Depends on the **order of rule applications**:

  **Example**: $D = \{p(a)\}$ and $\Sigma = \{p(x) \rightarrow \exists y \ q(y); \ p(x) \rightarrow q(x)\}$:
  
  Solution 1 = $\{p(a), q(u), q(a)\}$
  Solution 2 = $\{p(a), q(a)\}$

  $\Rightarrow$ Assume a canonical ordering.

- Can be **infinite**:

  **Example**: $D = \{p(a, b)\}$ and $\Sigma = \{p(x, y) \rightarrow \exists z \ p(y, z)\}$:
  
  Solution = $\{p(a, b), p(b, u_1), p(u_1, u_2), p(u_2, u_3), \ldots\}$

  $\Rightarrow$ Query answering for $D$ and TGDs alone is undecidable.

  $\Rightarrow$ Restrictions on TGDs and their interplay with EGDs.
Guarded and Linear Datalog+/-

A TGD $\sigma$ is guarded iff it contains an atom in its body that contains all universally quantified variables of $\sigma$.

Example:

- $r(X, Y), s(Y, X, Z) \rightarrow \exists W\ s(Z, X, W)$ is guarded, where $s(Y, X, Z)$ is the guard, and $r(X, Y)$ is a side atom;
- $r(X, Y), r(Y, Z) \rightarrow r(X, Z)$ is not guarded.

A TGD is linear iff it contains only a singleton body atom.

Example:

- $\text{manager}(M) \rightarrow \exists P\ \text{directs}(M, P)$ is linear;
- $r(X, Y), s(Y, X, Z) \rightarrow \exists W\ s(Z, X, W)$ is not linear.
Markov Logic Networks

- We use Markov logic networks (MLNs) to represent uncertainty in Datalog+/-.

- MLNs combine classical Markov networks (a.k.a. Markov random fields) with first-order logic (FOL).

- We assume a set of random variables \( X = \{X_1, \ldots, X_n\} \), where each \( X_i \) can take values in \( \text{Dom}(X_i) \).

- A value for \( X \) is a mapping \( x : X \rightarrow \bigcup_{i=1}^{n} \text{Dom}(X_i) \) such that \( x(X_i) \in \text{Dom}(X_i) \).

- MLN: set of pairs \((F, w)\), where \( F \) is a FO formula, and \( w \) is a real number.
The probability distribution represented by the MLN is:

\[ P(X = x) = \frac{1}{Z} \cdot \exp(\sum_j w_j \cdot n_j(x)), \]

where \( n_j \) is the number of ground instances of formula \( F_j \) made true by \( x \), \( w_j \) is the weight of formula \( F_j \), and \( Z = \sum_{x \in X} \exp(\sum_j w_j \cdot n_j(x)) \) (normalization constant).

**Exact inference** is \( \#P \)-complete, but MCMC methods obtain good approximations in practice.

A particularly costly step is the computation of \( Z \), but this is a one-time calculation.
Example

Consider the following MLN:

\[ \phi_1 : \text{ann}(S_1, l_1, \text{num}) \land \text{ann}(S_2, l_2, X) \land \text{overlap}(l_1, l_2) : 3 \]
\[ \phi_2 : \text{ann}(S_1, l_1, \text{shop}) \land \text{ann}(S_2, l_2, \text{mag}) \land \text{overlap}(l_1, l_2) : 1 \]
\[ \phi_3 : \text{ann}(S_1, l_1, \text{dl}) \land \text{ann}(S_2, l_2, \text{pers}) \land \text{overlap}(l_1, l_2) : 0.25 \]

Graph representation (for a specific set of constants):
Computing probabilities w.r.t. this MLN:

<table>
<thead>
<tr>
<th>λ₁</th>
<th>a₁</th>
<th>a₂</th>
<th>a₃</th>
<th>a₄</th>
<th>a₅</th>
<th>a₆</th>
<th>SAT</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>—</td>
<td>e⁰ / Z</td>
</tr>
<tr>
<td>2</td>
<td>False</td>
<td>False</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>φ₃</td>
<td>e⁰.²⁵ / Z</td>
</tr>
<tr>
<td>3</td>
<td>True</td>
<td>False</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>φ₁, φ₃</td>
<td>e³+⁰.²⁵ / Z</td>
</tr>
<tr>
<td>4</td>
<td>True</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>φ₁, φ₃</td>
<td>e³+⁰.²⁵ / Z</td>
</tr>
<tr>
<td>5</td>
<td>False</td>
<td>True</td>
<td>False</td>
<td>False</td>
<td>True</td>
<td>False</td>
<td>—</td>
<td>e⁰ / Z</td>
</tr>
<tr>
<td>6</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>φ₂</td>
<td>e¹ / Z</td>
</tr>
<tr>
<td>7</td>
<td>False</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>φ₂, φ₃</td>
<td>e¹+⁰.²⁵ / Z</td>
</tr>
<tr>
<td>8</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>φ₁, φ₂, φ₃</td>
<td>e³+¹+⁰.²⁵ / Z</td>
</tr>
</tbody>
</table>

... (64 possible settings for the binary random variables)
A probabilistic Datalog\+/– ontology consists of a classical Datalog\+/– ontology $O$ along with an MLN $M$.

Notation: $KB = (O, M)$

Formulas in $O$ are annotated with a set of pairs $\langle X_i = x_i \rangle$, with $x_i \in \{\text{true, false}\}$ (we also use 0 and 1, respectively). Variables that do not appear in the annotation are unconstrained.

Possible world: a set of pairs $\langle X_i = x_i \rangle$ where each $X_i \in X$ has a corresponding pair.

Basic intuition: given a possible world, a subset of the formulas in $O$ is induced.
Example Revisited

The following formulas were adapted from the previous examples to give rise to a probabilistic Datalog+/− ontology:

\[
\begin{align*}
\text{book}(X) & \rightarrow \text{editorialProd}(X) & : \{\} \\
\text{magazine}(X) & \rightarrow \text{editorialProd}(X) & : \{\} \\
\text{author}(X) & \rightarrow \text{person}(X,P) & : \{\} \\
\text{descLogic}(X) \land \text{author}(X) & \rightarrow \bot & : \{\text{ann}(X,I_1,dl) = 1 \land \text{ann}(X,I_2,pers) = 1 \land \text{overlap}(I_1,I_2) = 0\} \\
\text{shop}(X) \land \text{editorialProd}(X) & \rightarrow \bot & : \{\text{ann}(X,I_1,shop) = 1 \land \text{ann}(X,I_2,mag) = 1 \land \text{overlap}(I_1,I_2) = 0\} \\
\text{number}(X) \land \text{date}(X) & \rightarrow \bot & : \{\text{ann}(X,I_1,num) = 1 \land \text{ann}(X,I_1,date) = 1 \land \text{overlap}(I_1,I_2) = 0\}
\end{align*}
\]

Formulas with an empty annotation always hold.
Ranking Queries

- **Ranking Query (RQ)**: what are the ground atoms inferred from a KB, in decreasing order of probability?

- **Semantics**: the probability that a ground atom \( a \) is true is equal to the sum of the probabilities of possible worlds where the resulting KB entails the CQ \( a \).

- Recall that possible worlds are disjoint events.

- Unfortunately, computing probabilities of atoms is intractable:
  
  Theorem: Computing \( Pr(a) \) w.r.t. a given probabilistic ontology is \#P-hard in the data complexity.

- We now explore ways to tackle this uncertainty.
Conjunctive MLNs

• First, we propose a special class of MLNs:

A conjunctive MLN (cMLN) is an MLN in which all formulas \((F, w)\) in the set are such that \(F\) is a conjunction of atoms.

• This restriction allows us to define equivalence classes over the set of possible worlds w.r.t. \(M\):

  • Informally, two worlds are equivalent iff they satisfy the same formulas in \(M\).

  • Though there are still an exponential number of classes, there are some properties that we can leverage.

• Proposition 1: Given cMLN \(M\), deciding if an equivalence class \(C\) is empty is in PTIME.
Conjunctive MLNs: Properties

• Proposition 2: Given cMLN $\mathcal{M}$, and equivalence class $\mathcal{C}$, all elements in $\mathcal{C}$ can be obtained in linear time w.r.t. the size of the output.

• Proposition 3: Given cMLN $\mathcal{M}$, and worlds $\lambda_1$ and $\lambda_2$, we have that if $\lambda_1 \sim_\mathcal{M} \lambda_2$ then $Pr(\lambda_1) = Pr(\lambda_2)$.

• Proposition 4: Given cMLN $\mathcal{M}$, and worlds $\lambda_1$ and $\lambda_2$, deciding if $Pr(\lambda_1) \leq Pr(\lambda_2)$ is in PTIME.

• Computing exact probabilities in cMLNs, however, remains intractable:

  Theorem: Let $a$ be an atom; deciding if $Pr(a) \geq k$ is PP-hard in the data complexity.
• Also studying other kinds of probabilistic queries:
  ○ **Threshold** queries: what is the set of atoms that are inferred with probability at least $p$?
  ○ **Conjunctive** queries: what is the probability with which a conjunction of atoms is inferred?

• We are studying the **tractability** of all three kinds of queries under both sampling techniques.

• Also considering different kinds of **restrictions** on MLNs.
Summary of approximation and special-case algorithms:

<table>
<thead>
<tr>
<th>Problem</th>
<th>Monte Carlo Sampling</th>
<th>Top-down Enumeration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>General MLNs: Tractable, but no sound/complete guarantees TPM KBs: Bounded error and partial rankings can be guaranteed</td>
<td>cMLNs: Error is bounded and partial rankings guaranteed TPM KBs: Bounded error and partial rankings can be guaranteed</td>
</tr>
<tr>
<td>Threshold</td>
<td>General MLNs: #\textit{P}-Hard TPM KBs: Sound, complete under certain conditions</td>
<td>cMLNs: Sound and complete under certain conditions TPM KBs: Sound and complete under certain conditions</td>
</tr>
<tr>
<td>CQs</td>
<td>General MLNs: #\textit{P}-Hard TPM KBs: Sound</td>
<td>cMLNs: #\textit{P}-Hard TPM KBs: Tightest possible interval is guaranteed</td>
</tr>
</tbody>
</table>
Summary

- Presented an extension of the Datalog+/- family of languages with probabilistic uncertainty.
- Uncertainty in rules is expressed by means of annotations that refer to an underlying Markov Logic Network.
- The goal is to develop a language and algorithms capable of managing uncertainty in a principled and scalable way.

**Scalability** in our framework rests on two pillars:

- We combine scalable rule-based approaches from the DB literature with annotations reflecting uncertainty;
- Many possibilities for heuristic algorithms; MLNs are flexible, and sampling techniques may be leveraged.
References


## Outline

1. **Probabilistic Description Logics**
   - Motivation
   - Probabilistic Logics
   - \( \text{P-}SHIF(D) \) and \( \text{P-SHOIN}(D) \)

2. **Probabilistic Datalog+/−**
   - Datalog+/−
   - Markov Logic Networks
   - Probabilistic Datalog+/−

3. **Probabilistic Ontological Data Exchange**
   - Motivation and Overview
   - (Probabilistic) Ontological Data Exchange
   - Complexity Results
Motivation

Probabilistic ontological data exchange

- Ontological data exchange for integrated query answering over distributed ontologies on the Semantic Web.
- Ontological data exchange extending distributed ontology-based data access (OBDA).

Probabilities

- Automatically gathered and processed data (e.g., via information extraction, financial risk assessment) ⇒ probabilistic databases
- Uncertainty about the proper correspondence between items in distributed databases and ontologies (e.g., due to automatic generation) ⇒ probabilistic mappings
Overview

Probabilistic data exchange:

\[ D_{s1} \xrightarrow{\Sigma_{st1}} D_t \]

\[ \Sigma_{st1} \cup \Sigma_t : \text{TGDs from WA} \]
Overview

Probabilistic data exchange:

\[ D_{s1} \xrightarrow{\Sigma_{st_1}} D_{t1} \]

\[ \ldots \]

\[ D_{sn} \xrightarrow{\Sigma_{st_n}} D_{tm} \]

\[ \Sigma_{st} \cup \Sigma_t : \text{TGDs from WA} \]
Overview

Probabilistic data exchange:

\[ D_{s1} \xrightarrow{\Sigma_{st_1}} D_{t1} \]
\[ D_{sn} \xrightarrow{\Sigma_{st_n}} D_{tm} \]

\[ \Sigma_{st} \cup \Sigma_t : \text{TGDs from WA} \]
Probabilistic ontological data exchange: (PODE)

\[ \Sigma_s \cup \Sigma_{st} \cup \Sigma_t : \]
NCs and TGDs from WA, A, G, WG, S, WS, L, F, LF, AF, SF, GF
Probabilistic ontological data exchange: (PODE)

\[ \Sigma_s \cup \Sigma_{st} \cup \Sigma_t : \]
NCs and TGDs from WA, A, G, WG, S, WS, L, F, LF, AF, SF, GF
Probabilistic Databases

Probabilistic databases/instances:

- A probabilistic database (resp., probabilistic instance) is a probability space $Pr = (\mathcal{I}, \mu)$ such that $\mathcal{I}$ is the set of all databases (resp., instances) over a schema $\mathbf{S}$, and $\mu: \mathcal{I} \rightarrow [0, 1]$ is a function that satisfies $\sum_{I \in \mathcal{I}} \mu(I) = 1$.

Example:

<table>
<thead>
<tr>
<th>Possible database facts</th>
<th>Probabilistic database $Pr = (\mathcal{I}, \mu)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_a$</td>
<td>Researcher(Alice, UniversityOfOxford)</td>
</tr>
<tr>
<td>$r_p$</td>
<td>Researcher(Paul, UniversityOfOxford)</td>
</tr>
<tr>
<td>$p_{aml}$</td>
<td>Publication(Alice, ML, JMLR)</td>
</tr>
<tr>
<td>$p_{adb}$</td>
<td>Publication(Alice, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pdb}$</td>
<td>Publication(Paul, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pai}$</td>
<td>Publication(Paul, AI, AIJ)</td>
</tr>
</tbody>
</table>


Compact Encoding of Probabilistic Databases

Annotations and annotated atoms:

- **Elementary events** $e_i$: $e_1, \ldots, e_n$ with $n \geq 1$.
- **World** $w$: conjunctions $\ell_1 \land \cdots \land \ell_n$ of literals $\ell_i \in \{e_i, \neg e_i\}$.
- **Annotations** $\lambda$: Boolean combinations of elementary events:
  - each $e_i$ is an annotation $\lambda$;
  - if $\lambda_1$ and $\lambda_2$ are annotations, then also $\neg \lambda_1$ and $\lambda_1 \land \lambda_2$.
- **Annotated atoms** $a$: $\lambda$: atoms $a$ and annotations $\lambda$.

Uncertainty model:

- Bayesian network over $n$ binary random variables $E_1, \ldots, E_n$ with the domains $\text{dom}(E_i) = \{e_i, \neg e_i\}$. 

A set \( A \) of annotated atoms \( \{a_1 : \lambda_1, \ldots, a_l : \lambda_l\} \) along with a Bayesian network \( B \) compactly encodes a probabilistic database \( Pr = (I, \mu) \):

1. probability \( \mu(\lambda) \), for every annotation \( \lambda \): sum of the probabilities of all worlds in \( B \) in which \( \lambda \) is true;

2. probability \( \mu(D) \), for every database \( D = \{a_1, \ldots, a_m\} \in I \): probability of the conjunction \( \lambda = \lambda_1 \land \cdots \land \lambda_m \) of the annotations of its atoms. (Note that \( D \) is maximal with \( \lambda \).)
## Probabilistic Description Logics

**Probabilistic Datalog+/–**

**Probabilistic Ontological Data Exchange**

---

## Compact Encoding of Probabilistic Databases

### Example:

Possible database facts and their encoding

<table>
<thead>
<tr>
<th>Database Fact</th>
<th>Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_a$ Researcher(Alice, UniversityOfOxford)</td>
<td>true</td>
</tr>
<tr>
<td>$r_p$ Researcher(Paul, UniversityOfOxford)</td>
<td>$e_1 \vee e_2 \vee e_3 \vee e_4$</td>
</tr>
<tr>
<td>$p_{aml}$ Publication(Alice, ML, JMLR)</td>
<td>$e_1 \vee e_2$</td>
</tr>
<tr>
<td>$p_{adb}$ Publication(Alice, DB, TODS)</td>
<td>$\neg e_1 \land \neg e_2$</td>
</tr>
<tr>
<td>$p_{pdb}$ Publication(Paul, DB, TODS)</td>
<td>$e_1 \vee (\neg e_2 \land \neg e_3 \land e_4)$</td>
</tr>
<tr>
<td>$p_{pai}$ Publication(Paul, AI, AIJ)</td>
<td>$(\neg e_1 \land e_2) \lor (\neg e_1 \land e_3)$</td>
</tr>
</tbody>
</table>

Probabilistic database $Pr = (I, \mu)$

<table>
<thead>
<tr>
<th>$I_1$</th>
<th>$0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_2$</td>
<td>$0.2$</td>
</tr>
<tr>
<td>$I_3$</td>
<td>$0.15$</td>
</tr>
<tr>
<td>$I_4$</td>
<td>$0.075$</td>
</tr>
<tr>
<td>$I_5$</td>
<td>$0.075$</td>
</tr>
</tbody>
</table>

---

![Diagram of probabilistic database encoding](image)
Ontological Data Exchange (Syntax)

Ontological data exchange (ODE) problem \( \mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma_s, \Sigma_t, \Sigma_{st}) \):

- source schema \( S \),
- target schema \( T \), disjoint from \( S \),
- source ontology \( \Sigma_s \): finite set of TGDs and NCs over \( S \),
- target ontology \( \Sigma_t \): finite set of TGDs and NCs over \( T \),
- (source-to-target) mapping \( \Sigma_{st} \): finite set of TGDs and NCs over \( S \cup T \) with body(\( \sigma \)) and head(\( \sigma \)) over \( S \cup T \) and \( T \), resp..

Probabilistic ODE (PODE) problem \( \mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma_s, \Sigma_t, \Sigma_{st}, \mu_{st}) \):

- probabilistic (source-to-target) mapping \( \mu_{st} \): function \( \mu_{st} : 2^{\Sigma_{st}} \rightarrow [0, 1] \) such that \( \sum_{\Sigma' \subseteq \Sigma_{st}} \mu_{st}(\Sigma') = 1 \).
Ontological Data Exchange (Semantics)

- \(J\) is a solution (resp., universal solution) of \(I\) w.r.t. \(\Sigma\): 
  \(I \in \text{ins}(S), J \in \text{inst}(T), \) and \((I, J)\) is a model (resp., universal model) of \(\Sigma = \Sigma_s \cup \Sigma_t \cup \Sigma_{st}\)

- \(\text{Sol}_M\) (resp., \(\text{USol}_M\)): set of all pairs \((I, J)\) with \(J\) being a solution (resp., universal solution) for \(I\) w.r.t. \(\Sigma\)

- A probabilistic target instance \(Pr_t = (\mathcal{J}, \mu_t)\) is a probabilistic solution (resp., universal solution) for a probabilistic source database \(Pr_s = (\mathcal{I}, \mu_s)\) w.r.t. \(M = (S, T, \Sigma_s, \Sigma_t, \Sigma_{st})\) iff there exists a probability space \(Pr = (\mathcal{I} \times \mathcal{J}, \mu)\) such that:
  - The left and right marginals of \(Pr\) are \(Pr_s\) and \(Pr_t\), resp.: 
    - \(\sum_{J \in \mathcal{J}} (\mu(I, J)) = \mu_s(I)\) for all \(I \in \mathcal{I}\) and 
    - \(\sum_{I \in \mathcal{I}} (\mu(I, J)) = \mu_t(J)\) for all \(J \in \mathcal{J}\);
  - \(\mu(I, J) = 0\) for all \((I, J) \notin \text{Sol}_M\) (resp., \((I, J) \notin \text{USol}_M\)).
Probabilistic Description Logics
Probabilistic Datalog+–
Probabilistic Ontological Data Exchange

Ontological Data Exchange (Example)

- $\sigma_s : \text{Publication}(X, Y, Z) \rightarrow \text{ResearchArea}(X, Y)$
- $\sigma_{st} : \text{ResearchArea}(N, T) \land \text{Researcher}(N, U) \rightarrow \exists D \text{UResearchArea}(U, D, T)$
- $\sigma_t : \text{UResearchArea}(U, D, T) \rightarrow \exists Z \text{Lecturer}(T, Z)$

Possible source database facts

<table>
<thead>
<tr>
<th>$r_a$</th>
<th>Researcher(Alice, UoO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_p$</td>
<td>Researcher(Paul, UoO)</td>
</tr>
<tr>
<td>$p_{aml}$</td>
<td>Publication(Alice, ML, JMLR)</td>
</tr>
<tr>
<td>$p_{adb}$</td>
<td>Publication(Alice, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pdb}$</td>
<td>Publication(Paul, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pai}$</td>
<td>Publication(Paul, AI, AIJ)</td>
</tr>
</tbody>
</table>

Probabilistic source instance $P_{S} = (I, \mu_{S})$

| $I_1$ | $\{r_a, r_p, p_{aml}, p_{pdb}, r_{am}, r_{pdb}\}$ | 0.5 |
| $I_2$ | $\{r_a, r_p, p_{aml}, p_{pai}, r_{am}, r_{pai}\}$ | 0.2 |
| $I_3$ | $\{r_a, r_p, p_{adb}, p_{rai}, r_{adb}, r_{rai}\}$ | 0.15 |
| $I_4$ | $\{r_a, r_p, p_{adb}, p_{pdb}, r_{adb}, r_{pdb}\}$ | 0.075 |
| $I_5$ | $\{r_a, p_{adb}, r_{adb}\}$ | 0.075 |

Possible target instance facts

| $u_{ml}$ | UResearchArea(UoO, $N_1$, ML) |
| $u_{ai}$ | UResearchArea(UoO, $N_2$, AI) |
| $u_{db}$ | UResearchArea(UoO, $N_3$, DB) |
| $l_{ml}$ | Lecturer(ML, $N_4$) |
| $l_{ai}$ | Lecturer(AI, $N_5$) |
| $l_{db}$ | Lecturer(DB, $N_6$) |

Probabilistic universal solution $P_{T} = (J, \mu_{T})$

| $J_1$ | $\{u_{ml}, u_{db}\}$ | 0.5 |
| $J_2$ | $\{u_{ml}, u_{ai}\}$ | 0.2 |
| $J_3$ | $\{u_{ai}, u_{db}\}$ | 0.15 |
| $J_4$ | $\{u_{db}\}$ | 0.15 |
Ontological Data Exchange (Example)

- $\sigma_s : \text{Publication}(X, Y, Z) \rightarrow \text{ResearchArea}(X, Y)$

- $\sigma_{st} : \text{ResearchArea}(N, T) \land \text{Researcher}(N, U) \rightarrow \exists D \ U\text{ResearchArea}(U, D, T)$

- $\sigma_t : U\text{ResearchArea}(U, D, T) \rightarrow \exists Z \text{Lecturer}(T, Z)$

### Possible source database facts

| $r_a$ | Researcher(Alice, UoO) |
| $r_p$ | Researcher(Paul, UoO) |
| $p_{aml}$ | Publication(Alice, ML, JMLR) |
| $p_{adb}$ | Publication(Alice, DB, TODS) |
| $p_{pdb}$ | Publication(Paul, DB, TODS) |
| $p_{pai}$ | Publication(Paul, AI, AIJ) |

### Probabilistic source instance $Pr_S = (I, \mu_S)$

| $I_1 = \{r_a, r_p, p_{aml}, p_{pdb}, r_{aml}, r_{pdb}\}$ | 0.5 |
| $I_2 = \{r_a, r_p, p_{aml}, p_{pai}, r_{aml}, r_{pai}\}$ | 0.2 |
| $I_3 = \{r_a, r_p, p_{adb}, p_{pai}, r_{adb}, r_{pai}\}$ | 0.15 |
| $I_4 = \{r_a, r_p, p_{adb}, p_{pdb}, r_{adb}, r_{pdb}\}$ | 0.075 |
| $I_5 = \{r_a, p_{adb}, r_{adb}\}$ | 0.075 |

### Possible target instance facts

| $u_{ml}$ | UResearchArea(UoO, N_1, ML) |
| $u_{ai}$ | UResearchArea(UoO, N_2, AI) |
| $u_{db}$ | UResearchArea(UoO, N_3, DB) |
| $l_{ml}$ | Lecturer(ML, N_4) |
| $l_{ai}$ | Lecturer(AI, N_5) |
| $l_{db}$ | Lecturer(DB, N_6) |

### Probabilistic universal solution $Pr_t = (J, \mu_t)$

| $J_1 = \{u_{ml}, u_{db}, l_{ml}\}$ | 0.5 |
| $J_2 = \{u_{ml}, u_{ai}, l_{ml}\}$ | 0.2 |
| $J_3 = \{u_{ai}, u_{db}, l_{ai}\}$ | 0.15 |
| $J_4 = \{u_{db}, l_{db}\}$ | 0.15 |
Ontological Data Exchange (Example)

- \( \sigma_s : \text{Publication}(X, Y, Z) \rightarrow \text{ResearchArea}(X, Y) \)

- \( \sigma_{st} : \text{ResearchArea}(N, T) \land \text{Researcher}(N, U) \rightarrow \exists D \text{UResearchArea}(U, D, T) \)

- \( \sigma_t : \text{UResearchArea}(U, D, T) \rightarrow \exists Z \text{Lecturer}(T, Z) \)

Possible source database facts

<table>
<thead>
<tr>
<th>Fact</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ra</td>
<td>Researcher(Alice, UoO)</td>
</tr>
<tr>
<td>rp</td>
<td>Researcher(Paul, UoO)</td>
</tr>
<tr>
<td>paml</td>
<td>Publication(Alice, ML, JMLR)</td>
</tr>
<tr>
<td>pdb</td>
<td>Publication(Paul, DB, TODS)</td>
</tr>
<tr>
<td>ppai</td>
<td>Publication(Paul, AI, AIJ)</td>
</tr>
</tbody>
</table>

Probabilistic source instance \( Pr_S = (I, \mu_S) \)

<table>
<thead>
<tr>
<th>Instance</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I_1 ) = { ra, rp, paml, pdb, raaml, rapdb }</td>
<td>0.5</td>
</tr>
<tr>
<td>( I_2 ) = { ra, rp, paml, ppai, raaml, rapai }</td>
<td>0.2</td>
</tr>
<tr>
<td>( I_3 ) = { ra, rp, pdb, ppai, raadb, rapai }</td>
<td>0.15</td>
</tr>
<tr>
<td>( I_4 ) = { ra, rp, pdb, raadb, rapdb }</td>
<td>0.075</td>
</tr>
<tr>
<td>( I_5 ) = { ra, pdb, raadb }</td>
<td>0.075</td>
</tr>
</tbody>
</table>

Possible target instance facts

<table>
<thead>
<tr>
<th>Fact</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>uml</td>
<td>UResearchArea(UoO, N_1, ML)</td>
</tr>
<tr>
<td>uai</td>
<td>UResearchArea(UoO, N_2, AI)</td>
</tr>
<tr>
<td>uadb</td>
<td>UResearchArea(UoO, N_3, DB)</td>
</tr>
<tr>
<td>lml</td>
<td>Lecturer(ML, N_4)</td>
</tr>
<tr>
<td>lai</td>
<td>Lecturer(AI, N_5)</td>
</tr>
<tr>
<td>ldb</td>
<td>Lecturer(DB, N_6)</td>
</tr>
</tbody>
</table>

Probabilistic universal solution \( Pr_T = (J, \mu_T) \)

<table>
<thead>
<tr>
<th>Instance</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J_1 ) = { uml, uadb, lml, ldb }</td>
<td>0.5</td>
</tr>
<tr>
<td>( J_2 ) = { uml, uai, lml, lai }</td>
<td>0.2</td>
</tr>
<tr>
<td>( J_3 ) = { uai, uadb, lai, ldb }</td>
<td>0.15</td>
</tr>
<tr>
<td>( J_4 ) = { uadb, ldb }</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Ontological Data Exchange (Example)

- $\sigma_s : \text{Publication}(X, Y, Z) \rightarrow \text{ResearchArea}(X, Y)$
- $\sigma_{st} : \text{ResearchArea}(N, T) \land \text{Researcher}(N, U) \rightarrow \exists D \text{UResearchArea}(U, D, T)$
- $\sigma_t : \text{UResearchArea}(U, D, T) \rightarrow \exists Z \text{Lecturer}(T, Z)$

Possible source database facts

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_a$</td>
<td>Researcher(Alice, UoO)</td>
</tr>
<tr>
<td>$r_p$</td>
<td>Researcher(Paul, UoO)</td>
</tr>
<tr>
<td>$p_{aml}$</td>
<td>Publication(Alice, ML, JMLR)</td>
</tr>
<tr>
<td>$p_{adb}$</td>
<td>Publication(Alice, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pdb}$</td>
<td>Publication(Paul, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pai}$</td>
<td>Publication(Paul, AI, AIJ)</td>
</tr>
</tbody>
</table>

Probabilistic source instance $Pr_S = (I, \mu_S)$

- $I_1 = \{r_a, r_p, p_{aml}, p_{pdb}, r_{a_{aml}}, r_{a_{pdb}}\}$ with $\mu_I = 0.5$
- $I_2 = \{r_a, r_p, p_{aml}, p_{pai}, r_{a_{aml}}, r_{a_{pai}}\}$ with $\mu_I = 0.2$
- $I_3 = \{r_a, r_p, p_{adb}, p_{pai}, r_{a_{adb}}, r_{a_{pai}}\}$ with $\mu_I = 0.15$
- $I_4 = \{r_a, r_p, p_{adb}, p_{pdb}, r_{a_{adb}}, r_{a_{pdb}}\}$ with $\mu_I = 0.075$
- $I_5 = \{r_a, p_{adb}, r_{a_{adb}}\}$ with $\mu_I = 0.075$

Possible target instance facts

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{ml}$</td>
<td>UResearchArea(UoO, N_1, ML)</td>
</tr>
<tr>
<td>$u_{ai}$</td>
<td>UResearchArea(UoO, N_2, AI)</td>
</tr>
<tr>
<td>$u_{db}$</td>
<td>UResearchArea(UoO, N_3, DB)</td>
</tr>
<tr>
<td>$l_{ml}$</td>
<td>Lecturer(ML, N_4)</td>
</tr>
<tr>
<td>$l_{ai}$</td>
<td>Lecturer(AI, N_5)</td>
</tr>
<tr>
<td>$l_{db}$</td>
<td>Lecturer(DB, N_6)</td>
</tr>
</tbody>
</table>

Probabilistic universal solution $Pr_T = (J, \mu_T)$

- $J_1 = \{u_{ml}, u_{db}, l_{ml}, l_{db}\}$ with $\mu_I = 0.5$
- $J_2 = \{u_{ml}, u_{ai}, l_{ml}, l_{ai}\}$ with $\mu_I = 0.2$
- $J_3 = \{u_{ai}, u_{db}, l_{ai}, l_{db}\}$ with $\mu_I = 0.15$
- $J_4 = \{u_{db}, l_{db}\}$ with $\mu_I = 0.15$
Ontological Data Exchange (Example)

- $\sigma_s : \text{Publication}(X, Y, Z) \rightarrow \text{ResearchArea}(X, Y)$
- $\sigma_{st} : \text{ResearchArea}(N, T) \land \text{Researcher}(N, U) \rightarrow \exists D \text{UResearchArea}(U, D, T)$
- $\sigma_t : \text{UResearchArea}(U, D, T) \rightarrow \exists Z \text{Lecturer}(T, Z)$

Possible source database facts

<table>
<thead>
<tr>
<th>$r_a$</th>
<th>Researcher(Alice, UoO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_p$</td>
<td>Researcher(Paul, UoO)</td>
</tr>
<tr>
<td>$p_{aml}$</td>
<td>Publication(Alice, ML, JMLR)</td>
</tr>
<tr>
<td>$p_{adb}$</td>
<td>Publication(Alice, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pdb}$</td>
<td>Publication(Paul, DB, TODS)</td>
</tr>
<tr>
<td>$p_{pai}$</td>
<td>Publication(Paul, AI, AIJ)</td>
</tr>
</tbody>
</table>

Possible target instance facts

<table>
<thead>
<tr>
<th>$u_{ml}$</th>
<th>UResearchArea(UoO, $N_1$, ML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{ai}$</td>
<td>UResearchArea(UoO, $N_2$, AI)</td>
</tr>
<tr>
<td>$u_{db}$</td>
<td>UResearchArea(UoO, $N_3$, AI)</td>
</tr>
<tr>
<td>$l_{ml}$</td>
<td>Lecturer(ML, $N_4$)</td>
</tr>
<tr>
<td>$l_{ai}$</td>
<td>Lecturer(AI, $N_5$)</td>
</tr>
<tr>
<td>$l_{db}$</td>
<td>Lecturer(DB, $N_6$)</td>
</tr>
</tbody>
</table>

Probabilistic source instance $P_{S} = (I, \mu_S)$

| $I_1$ = {$r_a, r_p, p_{aml}, p_{pdb}, r_{aml}, r_{pdb}$} | 0.5 |
| $I_2$ = {$r_a, r_p, p_{aml}, p_{pai}, r_{aml}, r_{pai}$} | 0.2 |
| $I_3$ = {$r_a, r_p, p_{adb}, p_{pai}, r_{adb}, r_{pai}$} | 0.15 |
| $I_4$ = {$r_a, r_p, p_{adb}, p_{pdb}, r_{adb}, r_{pdb}$} | 0.075 |
| $I_5$ = {$r_a, p_{adb}, r_{adb}$} | 0.075 |

Probabilistic universal solution $P_{t} = (J, \mu_t)$

| $J_1$ = {$u_{ml}, u_{db}, l_{ml}, l_{db}$} | 0.5 |
| $J_2$ = {$u_{ml}, u_{ai}, l_{ml}, l_{ai}$} | 0.2 |
| $J_3$ = {$u_{ai}, u_{db}, l_{ai}, l_{db}$} | 0.15 |
| $J_4$ = {$u_{db}, l_{db}$} | 0.15 |

$I_1$ --- $J_1$
$I_2$ --- $J_2$
$I_3$ --- $J_3$
$I_4$ --- $J_4$
$I_5$ --- $J_4$
UCQs

Given:

- ODE problem $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma_s, \Sigma_t, \Sigma_{st})$;
- probabilistic source database $Pr_s = (\mathcal{I}, \mu_s)$;
- UCQ $q(\mathbf{X}) = \bigvee_{i=1}^{k} \exists \mathbf{Y}_i \Phi_i(\mathbf{X}, \mathbf{Y}_i,)$ over target schema.

Then, confidence of a tuple:

- $Pr_t(q(\mathbf{t}))$ for $Pr_t = (\mathcal{J}, \mu_t)$: sum of all $\mu_t(J)$ such that $q(\mathbf{t})$ evaluates to true in the instance $J \in \mathcal{J}$;
- $conf_{q}(\mathbf{t})$: confidence of a tuple $\mathbf{t}$ for $q$ in $Pr_s$ relative to $\mathcal{M}$: infimum of $Pr_t(q(\mathbf{t}))$ subject to all probabilistic solutions $Pr_t$ for $Pr_s$ relative to $\mathcal{M}$.
UCQs (Example)

Possible target instance facts

<table>
<thead>
<tr>
<th>u_{ml}</th>
<th>UResearchArea(University of Oxford, N_1, ML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>u_{ai}</td>
<td>UResearchArea(University of Oxford, N_2, AI)</td>
</tr>
<tr>
<td>u_{db}</td>
<td>UResearchArea(University of Oxford, N_3, DB)</td>
</tr>
<tr>
<td>l_{ml}</td>
<td>Lecturer(ML, N_4)</td>
</tr>
<tr>
<td>l_{ai}</td>
<td>Lecturer(AI, N_5)</td>
</tr>
<tr>
<td>l_{db}</td>
<td>Lecturer(DB, N_6)</td>
</tr>
</tbody>
</table>

Probabilistic universal solution \( Pr_t = (J, \mu_t) \)

<table>
<thead>
<tr>
<th>( J )</th>
<th>( \mu_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J_1 ) = {u_{ml}, u_{db}, l_{ml}, l_{db}}</td>
<td>0.5</td>
</tr>
<tr>
<td>( J_2 ) = {u_{ml}, u_{ai}, l_{ml}, l_{ai}}</td>
<td>0.2</td>
</tr>
<tr>
<td>( J_3 ) = {u_{ai}, u_{db}, l_{ai}, l_{db}}</td>
<td>0.15</td>
</tr>
<tr>
<td>( J_4 ) = {u_{db}, l_{db}}</td>
<td>0.15</td>
</tr>
</tbody>
</table>

\( Pr = \{((I_1, J_1), .5), ((I_2, J_2), .2), ((I_3, J_3), .15), ((I_4, J_4), .075), ((I_5, J_4), .075)\} \)

A student wants to know whether she can study both machine learning and databases at the University of Oxford:

\( q() = \exists X, Y (\exists Z (Lecturer(AI, X) \land UResearchArea(UnivOx, Z, AI)) \lor \exists Z (Lecturer(ML, Y) \land UResearchArea(UnivOx, Z, ML))) \).

Then, \( q \) yields the probability 0.85.
Computational Problems

Consistency:

- Given a (P)ODE problem $\mathcal{M}$ and a probabilistic source database $Pr_s$, decide whether there exists a (universal) probabilistic solution for $Pr_s$ relative to $\mathcal{M}$.

Threshold UCQ answering:

- Given a (P)ODE problem $\mathcal{M}$, a probabilistic source database $Pr_s$, a UCQ $q(X)$, a tuple $t$ of constants, and $\theta > 0$, decide whether $\text{conf}_Q(t) \geq \theta$ in $Pr_s$ w.r.t. $\mathcal{M}$.
Computational Problems

Classes of existential rules:

- linear full (LF), guarded full (GF), acyclic full (AF), sticky full (SF), full (F)
- acyclic (A), weakly acyclic (WA)
- linear (L), guarded (G), weakly guarded (WG)
- sticky (S), weakly sticky (WS)

Types of complexity:

- data complexity,
- fixed-program combined (fp-combined) complexity,
- bounded-arity combined (ba-combined) complexity,
- combined complexity
Relationships between Classes of Existential Rules
Complexity Results: Data Complexity

Data complexity of standard BCQ answering and consistency

<table>
<thead>
<tr>
<th>BCQs</th>
<th>consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>L, LF, AF</td>
<td>in $AC^0$</td>
</tr>
<tr>
<td>G</td>
<td>P</td>
</tr>
<tr>
<td>WG</td>
<td>EXP</td>
</tr>
<tr>
<td>S, SF</td>
<td>in $AC^0$</td>
</tr>
<tr>
<td>F, GF</td>
<td>P</td>
</tr>
<tr>
<td>A</td>
<td>in $AC^0$</td>
</tr>
<tr>
<td>WS, WA</td>
<td>P</td>
</tr>
</tbody>
</table>

$Pr_S$ with a polytree as BN $\Rightarrow$ consistency is in P in the data complexity for languages with BCQ answering in $AC^0$
### Complexity Results: fp-Combined Complexity

The fp-combined complexity of standard BCQ answering and consistency is as follows:

<table>
<thead>
<tr>
<th>BCQs</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>L, LF, AF</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>G</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>WG</td>
<td>EXP, EXP</td>
</tr>
<tr>
<td>S, SF</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>F, GF</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>A</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>WS, WA</td>
<td>NP, CONP</td>
</tr>
</tbody>
</table>

The diagram illustrates the complexity classes for different BCQs and consistency:

- WG
- F
- A
- S
- WS
- WA
- SF
- LF
- GF

The complexity classes are marked as follows:

- NP
- CONP
- EXP
- EXP
- NP
- CONP
- NP
- CONP
- NP
- CONP

The complexity classes NP and CONP indicate the difficulty of the problem, with NP being more complex than CONP.
### Complexity Results: ba-Combined Complexity

ba-combined complexity of standard BCQ answering and consistency

<table>
<thead>
<tr>
<th>BCQs</th>
<th>consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>L, LF, AF</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>G</td>
<td>EXP, EXP</td>
</tr>
<tr>
<td>WG</td>
<td>EXP, EXP</td>
</tr>
<tr>
<td>S, SF</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>F, GF</td>
<td>NP, CONP</td>
</tr>
<tr>
<td>A</td>
<td>NEXP, CONEXP</td>
</tr>
<tr>
<td>WS, WA</td>
<td>2EXP, 2EXP</td>
</tr>
</tbody>
</table>
### Complexity Results: Combined Complexity

combined complexity of standard BCQ answering and consistency

<table>
<thead>
<tr>
<th>BCQs</th>
<th>consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>L, LF, AF</td>
<td>PSPACE</td>
</tr>
<tr>
<td>G</td>
<td>2EXP</td>
</tr>
<tr>
<td>WG</td>
<td>2EXP</td>
</tr>
<tr>
<td>S, SF</td>
<td>EXP</td>
</tr>
<tr>
<td>F, GF</td>
<td>EXP</td>
</tr>
<tr>
<td>A</td>
<td>NEXP</td>
</tr>
<tr>
<td>WS, WA</td>
<td>2EXP</td>
</tr>
</tbody>
</table>
Summary of Complexity Results (Consistency)

Complexity of deciding the existence of a (universal) probabilistic solution (for both ODE and PODE problems):

<table>
<thead>
<tr>
<th>Data</th>
<th>(fp\text{-comb.})</th>
<th>(ba\text{-comb.})</th>
<th>Comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>L, LF, AF</td>
<td>CONP</td>
<td>CONP</td>
<td>CONP</td>
</tr>
<tr>
<td>G</td>
<td>CONP</td>
<td>CONP</td>
<td>EXP</td>
</tr>
<tr>
<td>WG</td>
<td>EXP</td>
<td>EXP</td>
<td>EXP</td>
</tr>
<tr>
<td>S, SF</td>
<td>CONP</td>
<td>CONP</td>
<td>CONP</td>
</tr>
<tr>
<td>F, GF</td>
<td>CONP</td>
<td>CONP</td>
<td>CONP</td>
</tr>
<tr>
<td>A</td>
<td>CONP</td>
<td>CONP</td>
<td>CONEXP</td>
</tr>
<tr>
<td>WS, WA</td>
<td>CONP</td>
<td>CONP</td>
<td>2EXP</td>
</tr>
</tbody>
</table>

All entries are completeness results; hardness holds even when any two variables are independent from each other.
Summary of Complexity Results (Threshold UCQ Entailment)

Complexity of deciding threshold query entailment (for both ODE and PODE problems; annotations are Boolean events under Bayesian networks).

<table>
<thead>
<tr>
<th>Data</th>
<th>fp-comb.</th>
<th>ba-comb.</th>
<th>Comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>L, LF, AF</td>
<td>PP</td>
<td>PP^{NP}</td>
<td>PP^{NP}</td>
</tr>
<tr>
<td>G</td>
<td>PP</td>
<td>PP^{NP}</td>
<td>EXP</td>
</tr>
<tr>
<td>WG</td>
<td>EXP</td>
<td>EXP</td>
<td>EXP</td>
</tr>
<tr>
<td>S, SF</td>
<td>PP</td>
<td>PP^{NP}</td>
<td>PP^{NP}</td>
</tr>
<tr>
<td>F, GF</td>
<td>PP</td>
<td>PP^{NP}</td>
<td>PP^{NP}</td>
</tr>
<tr>
<td>A</td>
<td>PP</td>
<td>PP^{NP}</td>
<td>NEXP</td>
</tr>
<tr>
<td>WS, WA</td>
<td>PP</td>
<td>PP^{NP}</td>
<td>2^{EXP}</td>
</tr>
</tbody>
</table>

All entries are completeness results; hardness holds even when any two variables are independent from each other.
Inconsistency-Tolerant Threshold UCQ Entailment

Repairing errors in probabilistic databases/instances; existential rules have no errors.

- repair of a deterministic database $D$ relative to $\Sigma$: maximal subset of $D$ that is consistent relative to $\Sigma$.

- repair of a probabilistic database $(\mathcal{I}, \mu)$ relative to $\Sigma$: consists of a repair of each $I \in \mathcal{I}$ with its probability $\mu(I)$

- $\text{conf}_q(t)$: confidence of a tuple $t$ for $q$ in $Pr_s$ relative to $\mathcal{M}$: infimum of $Pr_t(q(t))$ subject to all repairs of probabilistic solutions $Pr_t$ for $Pr_s$ relative to $\mathcal{M}$. 
Complexity Results (Inconsistency-Tolerant Threshold UCQ Entailment)

Consistency of deciding inconsistency-tolerant threshold query entailment (for both ODE and PODE problems; annotations are Boolean events under Bayesian networks).

<table>
<thead>
<tr>
<th>Data</th>
<th>fp-comb.</th>
<th>ba-comb.</th>
<th>Comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{\bot}$, $LF_{\bot}$, $AF_{\bot}$</td>
<td>$PP^{NP}$</td>
<td>$PP^{\Sigma_2^P}$</td>
<td>$PP^{\Sigma_2^P}$</td>
</tr>
<tr>
<td>$G_{\bot}$</td>
<td>$PP^{NP}$</td>
<td>$PP^{\Sigma_2^P}$</td>
<td>EXP</td>
</tr>
<tr>
<td>$WG_{\bot}$</td>
<td>EXP</td>
<td>EXP</td>
<td>EXP</td>
</tr>
<tr>
<td>$S_{\bot}$, $SF_{\bot}$</td>
<td>$PP^{NP}$</td>
<td>$PP^{\Sigma_2^P}$</td>
<td>$PP^{\Sigma_2^P}$</td>
</tr>
<tr>
<td>$F_{\bot}$, $GF_{\bot}$</td>
<td>$PP^{NP}$</td>
<td>$PP^{\Sigma_2^P}$</td>
<td>$PP^{\Sigma_2^P}$</td>
</tr>
<tr>
<td>$A_{\bot}$</td>
<td>$PP^{NP}$</td>
<td>$PP^{\Sigma_2^P}$</td>
<td>in $PP^{NEXP}$</td>
</tr>
<tr>
<td>$WS_{\bot}$, $WA_{\bot}$</td>
<td>$PP^{NP}$</td>
<td>$PP^{\Sigma_2^P}$</td>
<td>2EXP</td>
</tr>
</tbody>
</table>

All entries but the “in” ones are completeness results; hardness holds even when any two variables are independent from each other.
Summary

- ontological data exchange with probabilistic data
- ontological data exchange with probabilistic mappings
- compact encoding of probabilities via Boolean annotations under Bayesian networks as uncertainty models
- for the main classes of existential rules: data, fp-combined, ba-combined, and combined complexity for:
  - consistency
  - UCQ threshold entailment
  - inconsistency-tolerant UCQ threshold entailment
References
