Probabilistic Reasoning

Thomas Lukasiewicz

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

Outline

Outline

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 $\mathcal{SHIF}(\mathbf{D})$ and $\mathcal{SHOIN}(\mathbf{D})$ (which are the DLs behind OWL Lite and OWL DL, respectively).

Example

Description logic knowledge base L for an online store:

(1) $\mathit{Textbook} \sqsubseteq \mathit{Book}$; (2) $\mathit{PC} \sqcup \mathit{Laptop} \sqsubseteq \mathit{Electronics}$; $\mathit{PC} \sqsubseteq \neg \mathit{Laptop}$;
(3) $\mathit{Book} \sqcup \mathit{Electronics} \sqsubseteq \mathit{Product}$; $\mathit{Book} \sqsubseteq \neg \mathit{Electronics}$;
(4) $\mathit{Sale} \sqsubseteq \mathit{Product}$;
(5) $\mathit{Product} \sqsubseteq\ \geqslant 1\ \mathit{related}$; (6) $\geqslant 1\ \mathit{related} \sqcup \geqslant 1\ \mathit{related}^- \sqsubseteq \mathit{Product}$;
(7) $\mathit{related} \sqsubseteq \mathit{related}^-$; $\mathit{related}^- \sqsubseteq \mathit{related}$;
(8) $\mathit{Textbook}(\mathit{tb\_ai})$; $\mathit{Textbook}(\mathit{tb\_lp})$; (9) $\mathit{related}(\mathit{tb\_ai}, \mathit{tb\_lp})$;
(10) $\mathit{PC}(\mathit{pc\_ibm})$; $\mathit{PC}(\mathit{pc\_hp})$; (11) $\mathit{related}(\mathit{pc\_ibm}, \mathit{pc\_hp})$;
(12) $\mathit{provides}(\mathit{ibm}, \mathit{pc\_ibm})$; $\mathit{provides}(\mathit{hp}, \mathit{pc\_hp})$.

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 ϕ: Boolean combination of basic events
- Logical constraint $\psi \Leftarrow \phi$: events ψ and ϕ: "ϕ implies ψ".
- Conditional constraint $(\psi|\phi)[l, u]$: events ψ and ϕ, and $l, u \in [0, 1]$: "conditional probability of ψ given ϕ is in $[l, u]$".
- Probabilistic knowledge base $KB = (L, P)$:
  - finite set of logical constraints L,
  - finite set of conditional constraints P.

Example

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 Φ.
- $\mathcal{I}_\Phi$: all worlds for Φ.
- Probabilistic interpretation Pr: probability function on $\mathcal{I}_\Phi$.
- $\Pr(\phi)$: sum of all $\Pr(I)$ such that $I \in \mathcal{I}_\Phi$ and $I \models \phi$.
- $\Pr(\psi|\phi)$: if $\Pr(\phi) > 0$, then $\Pr(\psi|\phi) = \Pr(\psi \wedge \phi) \,/\, \Pr(\phi)$.
- Truth under Pr:
  - $\Pr \models \psi \Leftarrow \phi$ iff $\Pr(\psi \wedge \phi) = \Pr(\phi)$
    (iff $\Pr(\psi \Leftarrow \phi) = 1$).
  - $\Pr \models (\psi|\phi)[l, u]$ iff $\Pr(\psi \wedge \phi) \in [l, u] \cdot \Pr(\phi)$
    (iff either $\Pr(\phi) = 0$ or $\Pr(\psi|\phi) \in [l, u]$).

Example

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

	fly	$\neg$*fly*
bird	I_1	I_2
$\neg$*bird*	I_3	I_4

- Some probabilistic interpretations:

$\Pr_1$	*fly*	$\neg$*fly*
bird	19/40	1/40
$\neg$*bird*	10/40	10/40

$\Pr_2$	*fly*	$\neg$*fly*
bird	0	1/3
$\neg$*bird*	1/3	1/3

- $\Pr_1(fly \wedge bird) = 19/40$ and $\Pr_1(bird) = 20/40$.
- $\Pr_2(fly \wedge bird) = 0$ and $\Pr_2(bird) = 1/3$.
- $\neg fly \Leftarrow bird$ is false in $\Pr_1$, but true in $\Pr_2$.
- $(fly \mid bird)[.95, 1]$ 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.
- $KB \Vdash (\psi|\phi)[l, u]$: $(\psi|\phi)[l, u]$ is a logical consequence of KB iff every model of KB is also a model of $(\psi|\phi)[l, u]$.
- $KB \models_{tight} (\psi|\phi)[l, u]$: $(\psi|\phi)[l, u]$ 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:

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

- KB is satisfiable, since

  Pr with $\Pr(bird \wedge eagle \wedge have\_legs \wedge fly) = 1$ is a model.

- Some conclusions under logical entailment:

$$KB \mathrel{\|}\!\models (have\_legs \mid bird)[0.3,1], \quad KB \mathrel{\|}\!\models (fly \mid bird)[0.6,1].$$

- Tight conclusions under logical entailment:

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

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. ($\Rightarrow$ 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:

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

Tight conclusions under probabilistic lexicographic entailment:

$$KB \Vdash\!\sim^{lex}_{tight} (have\_legs \mid bird)[1,1], \quad KB \Vdash\!\sim^{lex}_{tight} (fly \mid bird)[0.95,1],$$
$$KB \Vdash\!\sim^{lex}_{tight} (have\_legs \mid eagle)[1,1], \quad KB \Vdash\!\sim^{lex}_{tight} (fly \mid eagle)[0.95,1].$$

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:

$$KB \Vdash\!\sim^{lex}_{tight} (have\_legs \mid bird)[1,1], \; KB \Vdash\!\sim^{lex}_{tight} (fly \mid bird)[1,1],$$
$$KB \Vdash\!\sim^{lex}_{tight} (have\_legs \mid penguin)[1,1], \; KB \Vdash\!\sim^{lex}_{tight} (fly \mid penguin)[0, 0.05].$$

Probabilistic knowledge base:

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

Tight conclusions under logical entailment:

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

Tight conclusions under probabilistic lexicographic entailment:

$$KB \Vdash\!\sim^{lex}_{tight} (have\_legs \mid bird)[0.99, 1],\ KB \Vdash\!\sim^{lex}_{tight} (fly \mid bird)[0.95, 1],$$
$$KB \Vdash\!\sim^{lex}_{tight} (have\_legs \mid penguin)[0.99, 1],\ KB \Vdash\!\sim^{lex}_{tight} (fly \mid penguin)[0, 0.05].$$

P-$\mathcal{SHIF}(\mathbf{D})$ and P-$\mathcal{SHOIN}(\mathbf{D})$: Key Ideas

- probabilistic generalization of the description logics $\mathcal{SHIF}(\mathbf{D})$ and $\mathcal{SHOIN}(\mathbf{D})$ 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:

$$
\begin{aligned}
(1)\ & \mathit{MalePacemakerPatient} \sqsubseteq \mathit{PacemakerPatient}, \\
& \mathit{FemalePacemakerPatient} \sqsubseteq \mathit{PacemakerPatient}, \\
(2)\ & \mathit{MalePacemakerPatient} \sqsubseteq \neg \mathit{FemalePacemakerPatient}, \\
(3)\ & \mathit{PacemakerPatient} \sqsubseteq \mathit{HeartPatient}, \\
(4)\ & \exists\, \mathit{HasIllnessSymptom}.\top \sqsubseteq \mathit{HeartPatient}, \\
& \exists\, \mathit{HasIllnessSymptom}^{-}.\top \sqsubseteq \mathit{IllnessSymptom}, \\
(5)\ & \mathit{HeartPatient}(\mathit{Tom}), \\
(6)\ & \mathit{MalePacemakerPatient}(\mathit{John}), \\
(7)\ & \mathit{FemalePacemakerPatient}(\mathit{Maria}), \\
(8)\ & \mathit{HasIllnessSymptom}(\mathit{John}, \mathit{Arrhythmia}), \\
& \mathit{HasIllnessSymptom}(\mathit{John}, \mathit{ChestPain}), \\
& \mathit{HasIllnessSymptom}(\mathit{John}, \mathit{BreathingDifficulties}), \\
& \mathit{HasIllnessStatus}(\mathit{John}, \mathit{Advanced}).
\end{aligned}
$$

Example

Terminological default and probabilistic knowledge:

$$
\begin{array}{ll}
(9) & (\mathit{HighBloodPressure} \mid \mathit{HeartPatient})[1,1], \\
(10) & (\neg \mathit{HighBloodPressure} \mid \mathit{PacemakerPatient})[1,1], \\
(11) & (\mathit{MalePacemakerPatient} \mid \mathit{PacemakerPatient})[0.4,1], \\
(12) & (\exists\, \mathit{HasHealthInsurance}.\mathit{PrivateHealthInsurance} \mid \mathit{HeartPatient})[0.9,1], \\
(13) & (\exists\, \mathit{HasIllnessSymptom}.\{\mathit{Arrhythmia}\} \mid \mathit{PacemakerPatient})[0.98,1], \\
 & (\exists\, \mathit{HasIllnessSymptom}.\{\mathit{ChestPain}\} \mid \mathit{PacemakerPatient})[0.9,1], \\
 & (\exists\, \mathit{HasIllnessSymptom}.\{\mathit{BreathingDifficulties}\} \mid \mathit{PacemakerPatient})[0.6,1].
\end{array}
$$

Example

Assertional probabilistic knowledge:

For individual Tom:

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

For individual Maria:

$$(15)\ (\exists\, \mathit{HasIllnessSymptom}.\{\mathit{BreathingDifficulties}\} \mid \top)[0.6, 1],$$
$$(16)\ (\exists\, \mathit{HasIllnessSymptom}.\{\mathit{ChestPain}\} \mid \top)[0.9, 1],$$
$$(17)\ (\exists\, \mathit{HasIllnessStatus}.\{\mathit{Final}\} \mid \top)[0.2, 0.8].$$

Complexity Results

	P-*DL-Lite*	P-$\mathcal{SHIF}$(**D**)	P-$\mathcal{SHOIN}$(**D**)
SAT	NP	EXP	NEXP
PTCON	NP	EXP	NEXP
PKBCON	NP	EXP	NEXP

	P-*DL-Lite*	P-$\mathcal{SHIF}$(**D**)	P-$\mathcal{SHOIN}$(**D**)
TLOGENT	$\mathrm{FP}^{\mathrm{NP}}$	FEXP	in $\mathrm{FP}^{\mathrm{NEXP}}$
TLEXENT	$\mathrm{FP}^{\mathrm{NP}}$	FEXP	in $\mathrm{FP}^{\mathrm{NEXP}}$

References

- T. Lukasiewicz. Expressive probabilistic description logics. *Artif. Intell.*, 172(6/7):852-883, 2008.

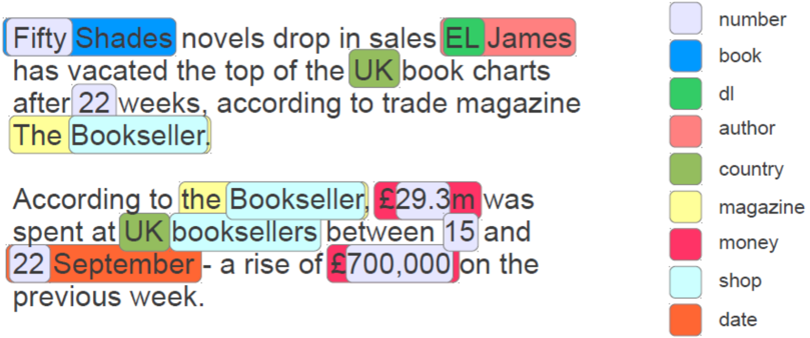
Outline

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).

Example

Consider the problem of entity extraction over the following text snippet:



Datalog+/$-$: Encoding Ontologies in Datalog

Plain Datalog allows for encoding some ontological axioms:

- concept inclusion axioms:

  $person(X) \leftarrow employee(X)$ iff $employee \sqsubseteq person$;

- role inclusion axioms:

  $manages(X, Y) \leftarrow reportsTo(Y, X)$ iff $reportsTo^{-1} \sqsubseteq manages$;

- concept and role membership axioms:

  $person(\text{John}) \leftarrow$ iff $person(\text{John})$;

  $manages(\text{Bill}, \text{John}) \leftarrow$ iff $manages(\text{Bill}, \text{John})$.

- transitivity axioms:

  $manages(X, Y) \leftarrow manages(X, Z), manages(Z, Y)$ iff (Trans $manages$)

However, it cannot express other important ontological axioms:

- concept inclusion axioms involving existential restrictions on roles in the head:

  $\mathit{Scientist} \sqsubseteq \exists \mathit{isAuthorOf}$;

- concept inclusion axioms stating concept disjointness:

  $\mathit{JournalPaper} \sqsubseteq \neg \mathit{ConferencePaper}$;

- functionality axioms:

  (funct $\mathit{hasFirstAuthor}$).

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):

  $\forall \mathbf{X} \forall \mathbf{Y} \exists \mathbf{Z}\ \Psi(\mathbf{X}, \mathbf{Z}) \leftarrow \Phi(\mathbf{X}, \mathbf{Y})$,
  where $\Phi(\mathbf{X}, \mathbf{Y})$ and $\Psi(\mathbf{X}, \mathbf{Z})$ are conjunctions of atoms;

  Example: $\exists P\, directs(M, P) \leftarrow manager(M)$;

- negative constraints:

  $\forall \mathbf{X} \perp \leftarrow \Phi(\mathbf{X})$,
  where $\Phi(\mathbf{X})$ is a conjunction of atoms;

  Example: $\perp \leftarrow c(X), c'(X)$;

- equality-generating dependencies (EGDs):

  $\forall \mathbf{X}\ X_i = X_j \leftarrow \Phi(\mathbf{X})$,
  where $X_i, X_j \in \mathbf{X}$, and $\Phi(\mathbf{X})$ is a conjunction of atoms

  Example: $Y = Z \leftarrow r_1(X, Y), r_2(Y, Z)$.

The Chase

Given:

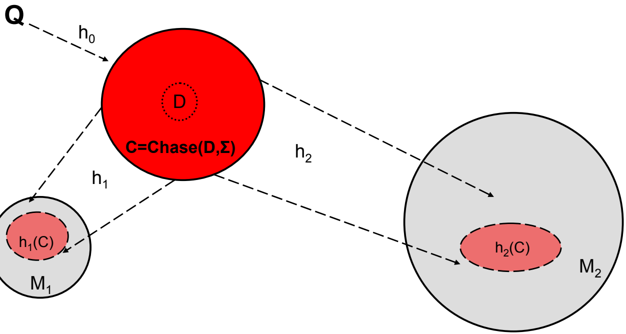
- D: database over $\mathrm{dom}(D)$.
- Σ: set of TGDs and/or EGDs

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 Σ 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 $\mathrm{dom}(D)$ elements in order to satisfy EGDs.

The Chase is a Universal Model



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

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

$$D \cup \Sigma \models Q \text{ iff } \text{chase}(D, \Sigma) \models Q$$

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 σ is guarded iff it contains an atom in its body that contains all universally quantified variables of σ.

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:

- $manager(M) \rightarrow \exists P\, directs(M, P)$ 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 $Dom(X_i)$.
- A value for X is a mapping $x\colon X \to \bigcup_{i=1}^{n} Dom(X_i)$ such that $x(X_i) \in Dom(X_i)$.
- 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.

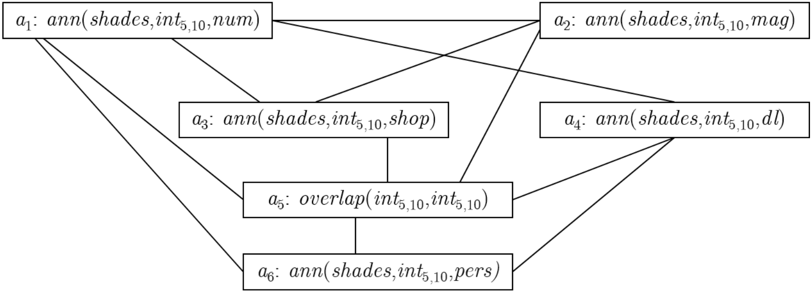
- A particularly costly step is the computation of Z, but this is a one-time calculation.

Example

Consider the following MLN:

$$\phi_1 : ann(S_1, I_1, num) \wedge ann(S_2, I_2, X) \wedge overlap(I_1, I_2) : 3$$
$$\phi_2 : ann(S_1, I_1, shop) \wedge ann(S_2, I_2, mag) \wedge overlap(I_1, I_2) : 1$$
$$\phi_3 : ann(S_1, I_1, dl) \wedge ann(S_2, I_2, pers) \wedge overlap(I_1, I_2) : 0.25$$

Graph representation (for a specific set of constants):



Computing probabilities w.r.t. this MLN:

λ_i	a_1	a_2	a_3	a_4	a_5	a_6	SAT	Probability
1	False	False	False	False	False	False	–	e^0 / Z
2	False	False	False	True	True	True	ϕ_3	$e^{0.25} / Z$
3	True	False	False	True	True	True	ϕ_1, ϕ_3	$e^{3+0.25} / Z$
4	True	False	True	True	True	True	ϕ_1, ϕ_3	$e^{3+0.25} / Z$
5	False	True	False	False	True	False	–	e^0 / Z
6	False	True	True	False	True	True	ϕ_2	e^1 / Z
7	False	True	True	True	True	True	ϕ_2, ϕ_3	$e^{1+0.25} / Z$
8	True	True	True	True	True	True	ϕ_1, ϕ_2, ϕ_3	$e^{3+1+0.25} / Z$

. . . (64 possible settings for the binary random variables)

Probabilistic Datalog+/– Ontologies

- 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 $x_i \in \{true, false\}$ (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:

$$book(X) \rightarrow editorialProd(X) \quad : \{\}$$

$$magazine(X) \rightarrow editorialProd(X) \quad : \{\}$$

$$author(X) \rightarrow person(X,P) \quad : \{\}$$

$$descLogic(X) \wedge author(X) \rightarrow \bot \quad : \{ann(X,I_1,dl) = 1 \wedge ann(X,I_2,pers) = 1 \; overlap(I_1,I_2) = 0\}$$

$$shop(X) \wedge editorialProd(X) \rightarrow \bot \quad : \{ann(X,I_1,shop) = 1 \wedge ann(X,I_2,mag) = 1 \; overlap(I_1,I_2) = 0\}$$

$$number(X) \wedge date(X) \rightarrow \bot \quad : \{ann(X,I_1,num) = 1 \wedge ann(X,I_1,date) = 1 \; overlap(I_1,I_2) = 0\}$$

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 M, and equivalence class C, all elements in C can be obtained in linear time w.r.t. the size of the output.
- Proposition 3: Given cMLN M, and worlds λ_1 and λ_2, we have that if $\lambda_1 \sim_M \lambda_2$ then $Pr(\lambda_1) = Pr(\lambda_2)$.
- Proposition 4: Given cMLN M, and worlds λ_1 and λ_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:

Problem	Monte Carlo Sampling	Top-down Enumeration
Ranking	General MLNs: Tractable, but no sound/complete guarantees TPM KBs: Bounded error and partial rankings can be guaranteed	cMLNs: Error is bounded and partial rankings guaranteed TPM KBs: Bounded error and partial rankings can be guaranteed
Threshold	General MLNs: *#P-Hard* TPM KBs: Sound, complete under certain conditions	cMLNs: Sound and complete under certain conditions TPM KBs: Sound and complete under certain conditions
CQs	General MLNs: *#P-Hard* TPM KBs: Sound	cMLNs: *#P-Hard* TPM KBs: Tightest possible interval is guaranteed

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

- T. Lukasiewicz, M. V. Martinez, G. Orsi, and G. I. Simari. Heuristic ranking in tightly coupled probabilistic description logics. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI 2012)*, pp. 554–563, 2012.

- G. Gottlob, T. Lukasiewicz, M. V. Martinez, and G. I. Simari. Query answering under probabilistic uncertainty in Datalog+/– ontologies. *Annals of Mathematics and Artificial Intelligence*, 69(1):37–72, Sept. 2013.

Outline

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

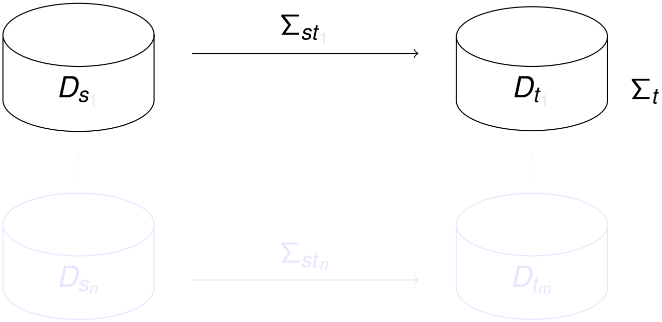
- Ontological data exchange for integrated query answering over distributed ontologies on the Semantic Web.
- Ontological data exchange extending distributed ontology-based data access (OBDA).

Probabilities

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

Overview

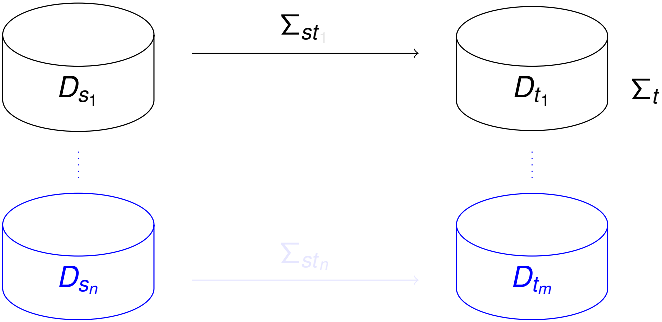
Probabilistic data exchange:



$\Sigma_{st} \cup \Sigma_t$: TGDs from WA

Overview

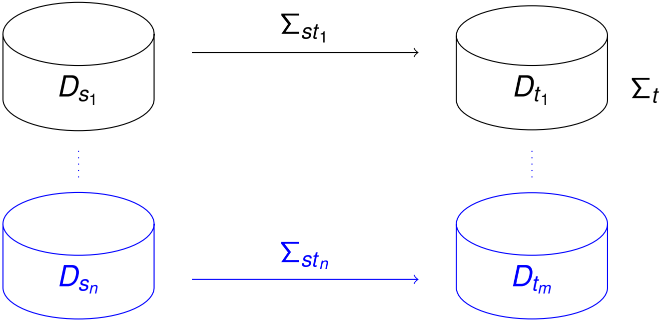
Probabilistic data exchange:



$\Sigma_{st} \cup \Sigma_t$: TGDs from WA

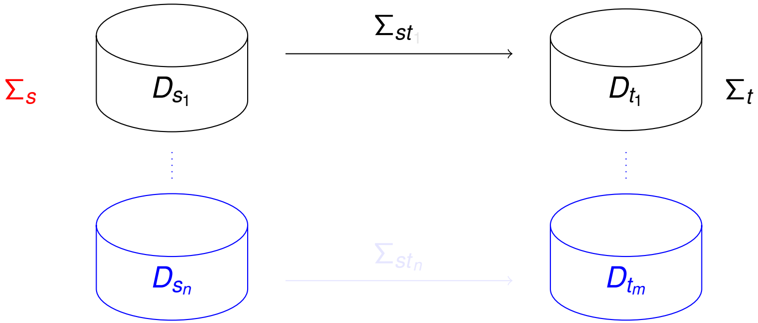
Overview

Probabilistic data exchange:



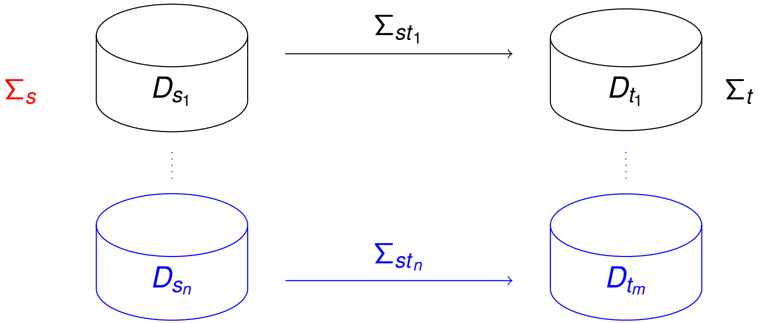
$\Sigma_{st} \cup \Sigma_t$: TGDs from WA

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 **S**, and $\mu \colon \mathcal{I} \to [0, 1]$ is a function that satisfies $\sum_{I \in \mathcal{I}} \mu(I) = 1$.

Example:

	Possible database facts
r_a	*Researcher*(Alice, UniversityOfOxford)
r_p	*Researcher*(Paul, UniversityOfOxford)
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 database $Pr = (\mathcal{I}, \mu)$	
$I_1 = \{\mathrm{r}_a, \mathrm{r}_p, \mathrm{p}_{aml}, \mathrm{p}_{pdb}\}$	0.5
$I_2 = \{\mathrm{r}_a, \mathrm{r}_p, \mathrm{p}_{aml}, \mathrm{p}_{pai}\}$	0.2
$I_3 = \{\mathrm{r}_a, \mathrm{r}_p, \mathrm{p}_{adb}, \mathrm{p}_{pai}\}$	0.15
$I_4 = \{\mathrm{r}_a, \mathrm{r}_p, \mathrm{p}_{adb}, \mathrm{p}_{pdb}\}$	0.075
$I_5 = \{\mathrm{r}_a, \mathrm{p}_{adb}\}$	0.075

Compact Encoding of Probabilistic Databases

Annotations and annotated atoms:

- Elementary events e_i: $e_1, \ldots, e_n$ with $n \geqslant 1$.
- World w: conjunctions $\ell_1 \wedge \cdots \wedge \ell_n$ of literals $\ell_i \in \{e_i, \neg e_i\}$.
- Annotations λ: Boolean combinations of elementary events:
  - each e_i is an annotation λ;
  - if λ_1 and λ_2 are annotations, then also $\neg\lambda_1$ and $\lambda_1 \wedge \lambda_2$.
- Annotated atoms $a\colon \lambda$: atoms a and annotations λ.

Uncertainty model:

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

Compact Encoding of Probabilistic Databases

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 $Pr = (\mathcal{I}, \mu)$:

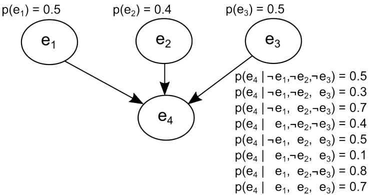
1. probability $\mu(\lambda)$, for every annotation λ: sum of the probabilities of all worlds in B in which λ is true;
2. probability $\mu(D)$, for every database $D = \{a_1, \ldots, a_m\} \in \mathcal{I}$: probability of the conjunction $\lambda = \lambda_1 \wedge \cdots \wedge \lambda_m$ of the annotations of its atoms. (Note that D is maximal with λ.)

Compact Encoding of Probabilistic Databases

Example:

Possible database facts and their encoding

r_a	*Researcher*(Alice, UniversityOfOxford)	true
r_p	*Researcher*(Paul, UniversityOfOxford)	$e_1 \vee e_2 \vee e_3 \vee e_4$
p_{aml}	*Publication*(Alice, ML, JMLR)	$e_1 \vee e_2$
p_{adb}	*Publication*(Alice, DB, TODS)	$\neg e_1 \wedge \neg e_2$
p_{pdb}	*Publication*(Paul, DB, TODS)	$e_1 \vee (\neg e_2 \wedge \neg e_3 \wedge e_4)$
p_{pai}	*Publication*(Paul, AI, AIJ)	$(\neg e_1 \wedge e_2) \vee (\neg e_1 \wedge e_3)$



Probabilistic database $Pr = (\mathcal{I}, \mu)$

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

Ontological Data Exchange (Syntax)

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

- source schema $\mathbf{S}$,
- target schema $\mathbf{T}$, disjoint from $\mathbf{S}$,
- source ontology Σ_s: finite set of TGDs and NCs over $\mathbf{S}$,
- target ontology Σ_t: finite set of TGDs and NCs over $\mathbf{T}$,
- (source-to-target) mapping Σ_{st}: finite set of TGDs and NCs over $\mathbf{S} \cup \mathbf{T}$ with *body*(σ) and *head*(σ) over $\mathbf{S} \cup \mathbf{T}$ and $\mathbf{T}$, resp..

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

- probabilistic (source-to-target) mapping μ_{st}: function $\mu_{st} \colon 2^{\Sigma_{st}} \to [0, 1]$ 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. Σ: $I \in \text{ins}(\mathbf{S})$, $J \in \text{inst}(\mathbf{T})$, and (I, J) is a model (resp., universal model) of $\Sigma = \Sigma_s \cup \Sigma_t \cup \Sigma_{st}$
- $Sol_{\mathcal{M}}$ (resp., $USol_{\mathcal{M}}$): set of all pairs (I, J) with J being a solution (resp., universal solution) for I w.r.t. Σ
- 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. $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma_s, \Sigma_t, \Sigma_{st})$ 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 $(I, J) \notin Sol_{\mathcal{M}}$ (resp., $(I, J) \notin USol_{\mathcal{M}}$).

Ontological Data Exchange (Example)

- σ_s : $\mathit{Publication}(X, Y, Z) \rightarrow \mathit{ResearchArea}(X, Y)$
- σ_{st} : $\mathit{ResearchArea}(N, T) \wedge \mathit{Researcher}(N, U) \rightarrow \exists D\ \mathit{UResearchArea}(U, D, T)$
- σ_t : $\mathit{UResearchArea}(U, D, T) \rightarrow \exists Z\ \mathit{Lecturer}(T, Z)$

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 $Pr_S = (\mathcal{I}, \mu_s)$

$I_1 = \{r_a, r_p, p_{aml}, p_{pdb}, ra_{aml}, ra_{pdb}\}$	0.5
$I_2 = \{r_a, r_p, p_{aml}, p_{pai}, ra_{aml}, ra_{pai}\}$	0.2
$I_3 = \{r_a, r_p, p_{adb}, p_{pai}, ra_{adb}, ra_{pai}\}$	0.15
$I_4 = \{r_a, r_p, p_{adb}, p_{pdb}, ra_{adb}, ra_{pdb}\}$	0.075
$I_5 = \{r_a, p_{adb}, ra_{adb}\}$	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 $Pr_t = (\mathcal{J}, \mu_t)$

$J_1 = \{u_{ml}, u_{db},$	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)

- σ_s : $\mathit{Publication}(X, Y, Z) \rightarrow \mathit{ResearchArea}(X, Y)$
- σ_{st} : $\mathit{ResearchArea}(N, T) \wedge \mathit{Researcher}(N, U) \rightarrow \exists D\ \mathit{UResearchArea}(U, D, T)$
- σ_t : $\mathit{UResearchArea}(U, D, T) \rightarrow \exists Z\ \mathit{Lecturer}(T, Z)$

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 $Pr_S = (\mathcal{I}, \mu_s)$

$I_1 = \{r_a, r_p, p_{aml}, p_{pdb}, ra_{aml}, ra_{pdb}\}$	0.5
$I_2 = \{r_a, r_p, p_{aml}, p_{pai}, ra_{aml}, ra_{pai}\}$	0.2
$I_3 = \{r_a, r_p, p_{adb}, p_{pai}, ra_{adb}, ra_{pai}\}$	0.15
$I_4 = \{r_a, r_p, p_{adb}, p_{pdb}, ra_{adb}, ra_{pdb}\}$	0.075
$I_5 = \{r_a, p_{adb}, ra_{adb}\}$	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 $Pr_t = (\mathcal{J}, \mu_t)$

$J_1 = \{u_{ml}, u_{db},$	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)

- σ_s : $\mathit{Publication}(X, Y, Z) \rightarrow \mathit{ResearchArea}(X, Y)$
- σ_{st} : $\mathit{ResearchArea}(N, T) \wedge \mathit{Researcher}(N, U) \rightarrow \exists D\ \mathit{UResearchArea}(U, D, T)$
- σ_t : $\mathit{UResearchArea}(U, D, T) \rightarrow \exists Z\ \mathit{Lecturer}(T, Z)$

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 $Pr_S = (\mathcal{I}, \mu_S)$

$I_1 = \{r_a, r_p, p_{aml}, p_{pdb}, ra_{aml}, ra_{pdb}\}$	0.5
$I_2 = \{r_a, r_p, p_{aml}, p_{pai}, ra_{aml}, ra_{pai}\}$	0.2
$I_3 = \{r_a, r_p, p_{adb}, p_{pai}, ra_{adb}, ra_{pai}\}$	0.15
$I_4 = \{r_a, r_p, p_{adb}, p_{pdb}, ra_{adb}, ra_{pdb}\}$	0.075
$I_5 = \{r_a, p_{adb}, ra_{adb}\}$	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 $Pr_t = (\mathcal{J}, \mu_t)$

$J_1 = \{u_{ml}, u_{db}, l_{ml}, l_{db}\}$	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

Ontological Data Exchange (Example)

- σ_s : $\mathit{Publication}(X, Y, Z) \rightarrow \mathit{ResearchArea}(X, Y)$
- σ_{st} : $\mathit{ResearchArea}(N, T) \wedge \mathit{Researcher}(N, U) \rightarrow \exists D\ \mathit{UResearchArea}(U, D, T)$
- σ_t : $\mathit{UResearchArea}(U, D, T) \rightarrow \exists Z\ \mathit{Lecturer}(T, Z)$

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 $Pr_S = (\mathcal{I}, \mu_s)$

$I_1 = \{r_a, r_p, p_{aml}, p_{pdb}, ra_{aml}, ra_{pdb}\}$	0.5
$I_2 = \{r_a, r_p, p_{aml}, p_{pai}, ra_{aml}, ra_{pai}\}$	0.2
$I_3 = \{r_a, r_p, p_{adb}, p_{pai}, ra_{adb}, ra_{pai}\}$	0.15
$I_4 = \{r_a, r_p, p_{adb}, p_{pdb}, ra_{adb}, ra_{pdb}\}$	0.075
$I_5 = \{r_a, p_{adb}, ra_{adb}\}$	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 $Pr_t = (\mathcal{J}, \mu_t)$

$J_1 = \{u_{ml}, u_{db}, l_{ml}, l_{db}\}$	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

Ontological Data Exchange (Example)

- σ_s : *Publication*(X, Y, Z) $\rightarrow$ *ResearchArea*(X, Y)
- σ_{st} : *ResearchArea*(N, T) $\wedge$ *Researcher*(N, U) $\rightarrow \exists$D *UResearchArea*(U, D, T)
- σ_t : *UResearchArea*(U, D, T) $\rightarrow \exists$Z *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 $Pr_S = (\mathcal{I}, \mu_s)$

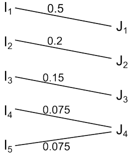
$I_1 = \{r_a, r_p, p_{aml}, p_{pdb}, ra_{aml}, ra_{pdb}\}$	0.5
$I_2 = \{r_a, r_p, p_{aml}, p_{pai}, ra_{aml}, ra_{pai}\}$	0.2
$I_3 = \{r_a, r_p, p_{adb}, p_{pai}, ra_{adb}, ra_{pai}\}$	0.15
$I_4 = \{r_a, r_p, p_{adb}, p_{pdb}, ra_{adb}, ra_{pdb}\}$	0.075
$I_5 = \{r_a, p_{adb}, ra_{adb}\}$	0.075

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 $Pr_t = (\mathcal{J}, \mu_t)$

$J_1 = \{u_{ml}, u_{db}, l_{ml}, l_{db}\}$	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



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
u_{ml}	*UResearchArea*(University of Oxford, N_1, ML)
u_{ai}	*UResearchArea*(University of Oxford, N_2, AI)
u_{db}	*UResearchArea*(University of Oxford, N_3, DB)
l_{ml}	*Lecturer*(ML, N_4)
l_{ai}	*Lecturer*(AI, N_5)
l_{db}	*Lecturer*(DB, N_6)

Probabilistic universal solution $Pr_t = (\mathcal{J}, \mu_t)$	
$J_1 = \{u_{ml}, u_{db}, l_{ml}, l_{db}\}$	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

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

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

$$\begin{aligned} q() = \exists X, Y(&\exists Z(Lecturer(AI, X) \wedge UResearchArea\,(UnivOx, Z, AI)) \\ &\vee \exists Z(Lecturer(ML, Y) \wedge UResearchArea(UnivOx, Z, ML))). \end{aligned}$$

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 $q(\mathbf{X})$, a tuple $\mathbf{t}$ of constants, and $\theta > 0$, decide whether $conf_Q(\mathbf{t}) \geqslant \theta$ in Pr_s w.r.t. $\mathcal{M}$.

Computational Problems

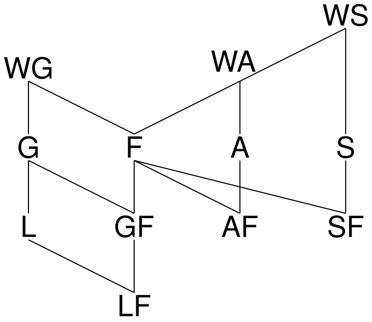
Classes of existential rules:

- 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

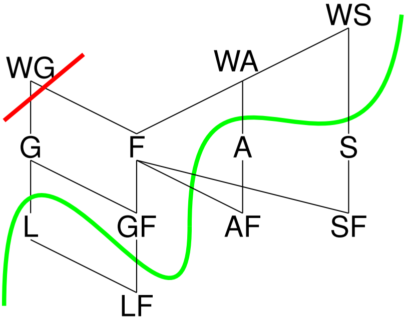
Relationships between Classes of Existential Rules



Complexity Results: Data Complexity

Data complexity of standard BCQ answering and consistency

	BCQs	consistency
L, LF, AF	in AC^0	CONP
G	P	CONP
WG	EXP	EXP
S, SF	in AC^0	CONP
F, GF	P	CONP
A	in AC^0	CONP
WS, WA	P	CONP

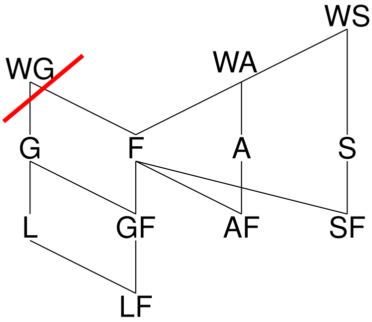


Pr_s with a polytree as BN $\Rightarrow$ consistency is in P in the data complexity for languages with BCQ answering in AC^0

Complexity Results: fp-Combined Complexity

fp-combined complexity of standard BCQ answering and consistency

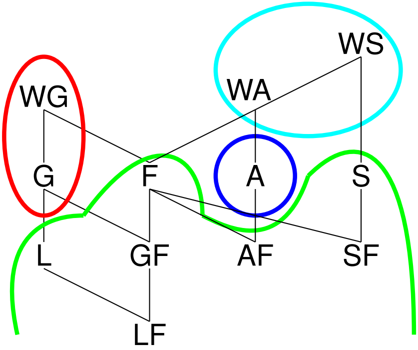
	BCQs	**consistency**
L, LF, AF	NP	coNP
G	NP	coNP
WG	EXP	EXP
S, SF	NP	coNP
F, GF	NP	coNP
A	NP	coNP
WS, WA	NP	coNP



Complexity Results: ba-Combined Complexity

ba-combined complexity of standard BCQ answering and consistency

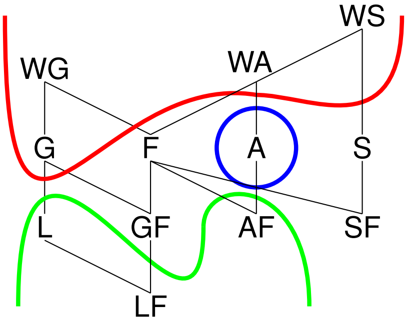
	BCQs	consistency
L, LF, AF	NP	coNP
G	EXP	EXP
WG	EXP	EXP
S, SF	NP	coNP
F, GF	NP	coNP
A	NEXP	coNEXP
WS, WA	2EXP	2EXP



Complexity Results: Combined Complexity

combined complexity of standard BCQ answering and consistency

	BCQs	consistency
L, LF, AF	PSPACE	PSPACE
G	2EXP	2EXP
WG	2EXP	2EXP
S, SF	EXP	EXP
F, GF	EXP	EXP
A	NEXP	coNEXP
WS, WA	2EXP	2EXP



Summary of Complexity Results (Consistency)

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

	Data	*fp*-**comb.**	*ba*-**comb.**	**Comb.**
L, LF, AF	coNP	coNP	coNP	PSPACE
G	coNP	coNP	EXP	2EXP
WG	EXP	EXP	EXP	2EXP
S, SF	coNP	coNP	coNP	EXP
F, GF	coNP	coNP	coNP	EXP
A	coNP	coNP	coNEXP	coNEXP
WS, WA	coNP	coNP	2EXP	2EXP

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).

	Data	***fp*-comb.**	***ba*-comb.**	**Comb.**
L, LF, AF	PP	$\mathrm{PP}^{\mathrm{NP}}$	$\mathrm{PP}^{\mathrm{NP}}$	PSPACE
G	PP	$\mathrm{PP}^{\mathrm{NP}}$	EXP	2EXP
WG	EXP	EXP	EXP	2EXP
S, SF	PP	$\mathrm{PP}^{\mathrm{NP}}$	$\mathrm{PP}^{\mathrm{NP}}$	EXP
F, GF	PP	$\mathrm{PP}^{\mathrm{NP}}$	$\mathrm{PP}^{\mathrm{NP}}$	EXP
A	PP	$\mathrm{PP}^{\mathrm{NP}}$	NEXP	NEXP
WS, WA	PP	$\mathrm{PP}^{\mathrm{NP}}$	2EXP	2EXP

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 Σ: maximal subset of D that is consistent relative to Σ.
- repair of a probabilistic database $(\mathcal{I}, \mu)$ relative to Σ: consists of a repair of each $I \in \mathcal{I}$ with its probability $\mu(I)$
- $conf_q(\mathbf{t})$: 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 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).

	Data	***fp*-comb.**	***ba*-comb.**	**Comb.**
$L_\perp$, $LF_\perp$, $AF_\perp$	PP^{NP}	$PP^{\Sigma_2^p}$	$PP^{\Sigma_2^p}$	PSPACE
$G_\perp$	PP^{NP}	$PP^{\Sigma_2^p}$	EXP	2EXP
$WG_\perp$	EXP	EXP	EXP	2EXP
$S_\perp$, $SF_\perp$	PP^{NP}	$PP^{\Sigma_2^p}$	$PP^{\Sigma_2^p}$	EXP
$F_\perp$, $GF_\perp$	PP^{NP}	$PP^{\Sigma_2^p}$	$PP^{\Sigma_2^p}$	EXP
$A_\perp$	PP^{NP}	$PP^{\Sigma_2^p}$	in PP^{NEXP}	in PP^{NEXP}
$WS_\perp$, $WA_\perp$	PP^{NP}	$PP^{\Sigma_2^p}$	2EXP	2EXP

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

- T. Lukasiewicz, M. V. Martinez, L. Predoiu, G. I. Simari. Existential rules and Bayesian networks for probabilistic ontological data exchange. Proc. RuleML 2015
- T. Lukasiewicz, M. V. Martinez, L. Predoiu, G. I. Simari. Basic probabilistic ontological data exchange with existential rules. Proc. AAAI 2016