Logical and Probabilistic Knowledge Representation and Reasoning

Leopoldo Bertossi

Carleton University

School of Computer Science

Ottawa, Canada

A Bit of Knowledge Representation

- KR has to do with: [36]
  - Representing knowledge in computers
  - Enable computers to "reason" with that knowledge
- Traditional areas of KR (among others):
  1. Automated diagnosis
  2. Reasoning in the presence of uncertainty [16]

     Knowledge that is ambiguous, incomplete, inconsistent, etc.
  3. Commonsense reasoning (may fall under 2.)
  4. Reasoning about causality (in particular, action/effect)
  5. Probabilistic reasoning
- 1.-4. not necessarily or intrinsically of probabilistic nature

A "Logical" Approach:

Example: Commonsense reasoning

(*) Knowledge: *"Tweety is a bird"*

Consequence: *"Tweety flies"*

- Possible interpretation of the use of (*): A "default rule" is applied

"In the absence of opposite information, conclude that the bird flies ..."

"In the absence of information that the bird is exceptional, conclude that the bird flies ..."

In more "logical" terms:

$$KB = \{\, \forall x(Bird(x) \land \neg Abnormal(x) \rightarrow Flies(x)),\;\; Bird(tweety)\,\}$$

"Since there is no evidence of Tweety's abnormality, conclude it flies"

The extension of predicate $Abnormal$ is minimized ...

Intuition formalized as "circumscription" (John McCarthy, 1980)

- A “non-monotonic”, non-classical logic ...
- Alternatively, “Default Logic” (Ray Reiter, 1980)

  Deduction rule: $$\frac{Bird(x)\ :\ Flies(x)}{Flies(x)}$$

“Given that x is bird, if it can be consistently assumed that x flies, conclude x flies”

Example: Diagnosis Reasoning

Knowledge: *Flu* $\Rightarrow$ *Fiver*

Fiver (evidence, observation)

Conclusion: *Flu*

- The diagnosis problem in AI has also been treated using logical methods [29]

  For example, through abduction:

Given a knowledge base K, and an observation O, we want an explanation E, expressed in logical terms, such that:

$$K \wedge E \Rightarrow O$$ (usually with an extra minimality assumption on E)

- Again, a non-monotonic, non-classical logic ...

A “Probabilistic” Approach:

- Problems 1.-4. may also have probabilistic elements, or be treated as such, even without explicit initial probabilities (cf. slide 2)

- For example, a default rule may be treated as a probabilistic or statistical statement (with high probability)

  And the consequences may be probabilistic too

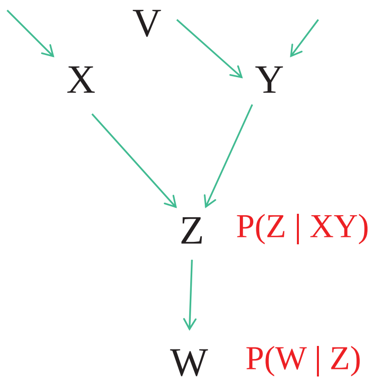
- Or as conditional probabilities: $P(\mathit{flies} \mid \mathit{bird}) = 0.95$

  “the probability of flying being a bird is 0.95”

- How to represent conditional probabilities?

- “Bayesian networks” appear [5]

  How to represent probabilistic knowledge; in particular, about the propagation of conditional probabilities?



The probabilistic model emerges from the network structure plus conditional probabilities, and other assumptions

- The model is a join probability distribution P over all variables
- The structure conveys causal elements
- Conditional probabilities are initially given, and may be subjective
- Additional assumptions ...

Of the kind: the conditional probability on a node given its parent nodes does not depend upon its non-descendant nodes (locally Markovian network)

- A full approach to uncertain and causal reasoning is based on Bayesian networks [26, 27]

- And Markov networks: as above, but non-directed

  Instead of conditional probabilities they represent initial local join probability distributions

  Cliques in the NW have associated "potential functions", which define join probability distributions (and factors of the global distribution)

- Using conditional probabilities allows attacking several KR problems

- In particular, maximum-likelihood reasoning

  For example, abductive diagnosis may also be probabilistic/statistical

Given a probability distribution, me may, e.g. find a maximum-likelihood explanation E:

$$\text{arg max}_E \; P(O \mid K \wedge E)$$

One that maximizes the probability of the observation, which was observed after all ...

Putting Things Together:

- Research that unifies or connects different probabilistic approaches are of crucial relevance
- Traditionally, the “logical-” and “probabilistic schools” have been separate and competitors
- In the last few years they have become complementary approaches
- Today, KR problems are attacked with mathematical models/techniques that involve simultaneously logic (rather classical) and probability

Logic $+$ Probability in AI

- Different forms of KR combine logic and probability

  Each of them with a domain of intended applications

  Just a couple of approaches, to give the idea ...

ME Distributions:

- Logical consequences from a knowledge base KB?
- In KB there is:
  - "hard" knowledge, e.g. $emu \rightarrow bird$
  - "soft", conditional, probabilistic rules, of the form r: $(\alpha|\beta)[p]$

    E.g. $r_v: \ (flies|bird)[0.9]$

- There is a collection $\mathcal{C}$ of worlds, $\mathcal{W}$, associated to KB that satisfy the hard knowledge, but not necessarily $\beta \to \alpha$

  (conditionals as classical implications)

- There are also several probability distributions (candidates) P over $\mathcal{C}$: $\mathcal{W} \mapsto P(\mathcal{W})$

  We restrict ourselves to the collection $\mathcal{P}$ of distributions P that satisfy the KB, in particular, the conditionals rules r: $(\alpha|\beta)[p]$

  For example, it must hold:[1] $P(\alpha|\beta) = p$ (and $P(\beta) > 0$)

- The logical-probabilistic consequences of KB are those with high probability

- But, under which of the distributions in $\mathcal{P}$?

  There may be some that are better than others ...

---

[1] $P(\alpha|\beta) := \frac{P(\alpha \wedge \beta)}{P(\beta)} := \frac{P(\{\mathcal{W} \mid \mathcal{W} \models \alpha \wedge \beta\})}{P(\{\mathcal{W} \mid \mathcal{W} \models \beta\})}$

- We choose one having maximum entropy: [17, 24]

$$P^\star \;:=\; \text{arg max}_{P\in\mathcal{P}} \; Entropy(P)$$

$$=\; \text{arg max}_{P\in\mathcal{P}} \; -\sum_{\mathcal{W}\in\mathcal{C}} P(\mathcal{W})\times ln(P(\mathcal{W}))$$

Distribution with less arbitrary assumptions/structure, maximum disorder, maximum independence, less information for free:

- The logical-probabilistic consequences of KB are those with high probability under $P^\star$

$$KB \;\models_{\text{l-p}}\; \varphi \;\;:\Longleftrightarrow\;\; P^\star(\varphi) \geq 1-\epsilon$$

- The computation of consequences from KB via the maximum entropy distribution can be automated [18]

- It is a philosophically interesting choice [25, 3]

Markov Logic Networks: (MLNs) [30, 10]

- MLNs combine FO logic and Markov Networks (MNs) in the same representation

- Coming from Bayesian, Belief and MNs, they are used not only for KR, but also in machine learning (ML)

  Networks can be learned from data for producing KR models and new inferences

- We have a knowledge base KB in FO logic

- The interpretations (models, worlds, ...) have associated probabilities

- There may be “models” that are more likely than others

- A world that violates a formula is not invalid (non-model), but only less probable

- Weights w_i are assigned (real-valued) to each of the F formulae F_i in KB

  The higher the weight, the higher is the difference between a world that satisfies the formula and another that doesn't (with everything else the same)

Formula	Weight
$\forall x : Steal(x) \Rightarrow Prison(x)$	3
$\forall x \forall y : CrimePartners(x, y) \wedge Steal(x) \Rightarrow Prison(y)$	1.5

- The weight of a formula captures the way the probability decreases when an "instance" of the formula is violated

- There is a fixed, finite, domain of constants, with which predicates and formulae can be instantiated, e.g. $Dom = \{bob, anna, ...\}$

  Producing atoms and instantiated formulae (of the propositional kind), e.g. $CrimePartners(bob, anna)$, $Steal(bob) \rightarrow Prison(bob)$, resp.

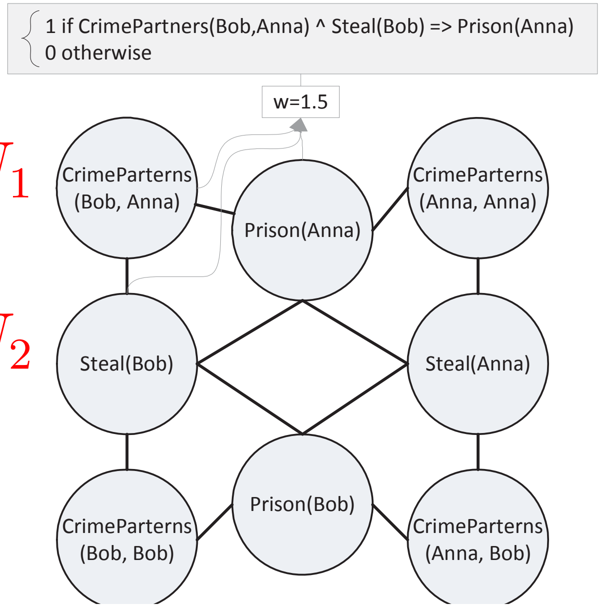
- A world is a set of atoms (a Herbrand structure), e.g.

$$\mathcal{W}_1 = \{CrimePartners(bob, anna), Steal(bob)\}$$

$$\mathcal{W}_2 = \{CrimePartners(bob, anna), Steal(bob), Prison(anna)\}$$

N_1

N_2



The instantiated atoms are the M nodes of the network (not directed)

Each node N takes the value $0/1$ depending on whether it is false or true (in a world $\mathcal{W}$)

Each node N is a Bernoulli random variable X^N

$\mathcal{W}_1$ represented by random vector: $\mathcal{X} = \langle X^{N_1}, X^{N_2}, X^{N_3}, \ldots \rangle$
$= \langle 1, 1, 0, \ldots, 0 \rangle$

- Edge between two nodes if they appear in a same instantiated formula

- As in ML, the representation has associated "features", which are measurable variables (functions, i.e. random variables)

- Each instantiation of a formula generates a "feature", with value 1 if true in the world $\mathcal{W}$, and 0, otherwise

- All these ingredients plus a log-linear "potential function" (to generate join probability distributions) give the probability of world $\mathcal{W}$ associated to $\mathbf{x} \in \{0,1\}^M$:

$$P(\mathcal{X} = \mathbf{x}) = \frac{1}{Z} exp(\sum_{i=1}^{F} w_i \times n_i(\mathbf{x}))$$

  $n_i(\mathbf{x})$: number of instantiations of F_i true in world $\mathbf{x}$

  Z normalizes considering all possible worlds:

$$Z = \sum_{\mathbf{z} \in \{0,1\}^M} exp(\sum_{i=1}^{F} w_i \times n_i(\mathbf{z}))$$

- This is just the start ...

- How to learn an MLN? (ML) [12, 20, 31, 21]

- How to do inference with MLNs?

Inference in Markov Logic Networks:

- Inference under MLNs is a probabilistic nature

  A basic inference task is computing the probability of a world, as in slide 15, but it could be more interesting: for a logical sentence φ

  $$P(\varphi) := P(\mathbf{x} \in \{0,1\}^M \mid \varphi \text{ is true in } \mathbf{x}) = \sum_{\mathbf{x} \in \{0,1\}^M \ \mid \ \varphi \cdots} P(\mathbf{x})$$

  If this probability is sufficiently high, we can accept it as an inference from the MLN

- Computing the probabilities amounts, directly or not, to counting models (possibly with specific properties) (see slide 15)

  And combining the resulting numbers with the weights; to use the probabilistic formulas above

- An active area of research in these days

  Different areas converge: model counting in logic (around SAT-related problems), graph theory, and data management [23, 2, 13]

- Instantiating and counting next may lead to combinatorial explosion

  Can we do better? Counting models without instantiating?

- Can we approximate model counting (and probabilities), also without instantiating?

- This is what is called "lifted" inference [19, 14, 15, 35]

  Lifted up to the FO representation ...

Some Lifted Inference: In the end, it attacks the problem of weighted (first-order) model counting [34]

- Probabilistic inference based on lifted weighted model counting runs in polynomial time in the number of objects in the domain

  $\longrightarrow$ data complexity!

  (Not necessarily in the number of predicates, formulas, logical variables)

- Lifted inference avoids grounding and counting by exploiting:

  - Stochastic independence

    Related to the (underlying) graph structure, in particular tree-width

    Markovian assumptions and conditional independence

- Symmetries, exchangeability [22]

  A set of random variables $\mathcal{X} = \{X^1, \ldots, X^n\}$ is fully exchangeable if and only if $P(X^1 = x_1, \ldots, X^n = x_n) = P(X^1 = x_{\pi(1)}, \ldots, X^n = x_{\pi(n)})$ for all permutations π of $\{1, \ldots, n\}$

  Best understood thinking of Bernoulli random variables indicating head/tail

  Exchangeability means that it is only the number of heads that matters and not their particular order

- Reasoning at FO level

- Reasoning about groups of objects

Example: (weighted FO model counting, not necessarily for MLNs)[2]

Smokes(Alice)	Smokes(Bob)	Friends(Alice,Bob)	Friends(Bob,Alice)
0	0	0	0
⋮	⋮	⋮	⋮
1	0	1	0
⋮	⋮	⋮	⋮
1	1	1	1

A schema

Possible worlds

Logical interpretations

A logical theory:

$$\forall xy(Smokes(x) \wedge Friends(x,y) \rightarrow Smokes(y))$$

Smokes(Alice)	Smokes(Bob)	Friends(Alice,Bob)	Friends(Bob,Alice)	theory
0	0	0	0	1
⋮	⋮	⋮	⋮	⋮
1	0	1	0	0
⋮	⋮	⋮	⋮	⋮
1	1	1	1	1

Models of the theory

$\Longleftarrow$ FO model count $\sim$ #SAT

[2] Borrowed from Guy van den Broeck

Smokes(Alice)	Smokes(Bob)	Friends(Alice,Bob)	Friends(Bob,Alice)	theory	weight
0	0	0	0	1	$2 \cdot 2 \cdot 1 \cdot 1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	1	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	1	1	$1 \cdot 1 \cdot 4 \cdot 4$

Now a logical theory and weights for predicates:

$Smokes \mapsto 1$

$\neg Smokes \mapsto 2$

$Friends \mapsto 4$

$\neg Friends \mapsto 1$

weighted FO model count $\sim$ partition function

Example:

1. Logical sentence: Domain:

$$\forall x(Stress(x) \rightarrow Smokes(x))$$

$Alice$

$Stress(Alice)$	$Smokes(Alice)$	Formula
0	0	1
0	1	1
1	0	0
1	1	1

3 models!

2. Logical sentence: Domain:

$$\forall x(Stress(x) \rightarrow Smokes(x))$$

n people

3^n models!

3. Logical sentence: Domain:

$\forall y(ParentOf(y) \wedge Female \rightarrow MotherOf(y))$ n people

- If $Female$: $\forall y(ParentOf(y) \rightarrow MotherOf(y))$
- If not $Female$: $true$

$(3^n + 4^n)$ models!

4. Logical sentence: Domain:

$\forall xy(ParentOf(x, y) \wedge Female(x) \rightarrow MotherOf(x, y))$ n people

$(3^n + 4^n)^n$ models!

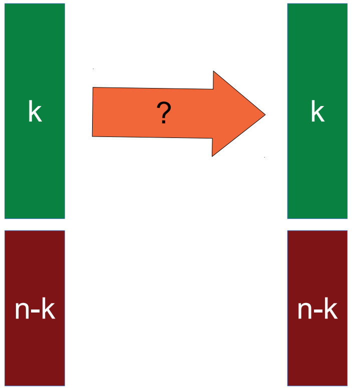
5. Logical sentence: Domain:

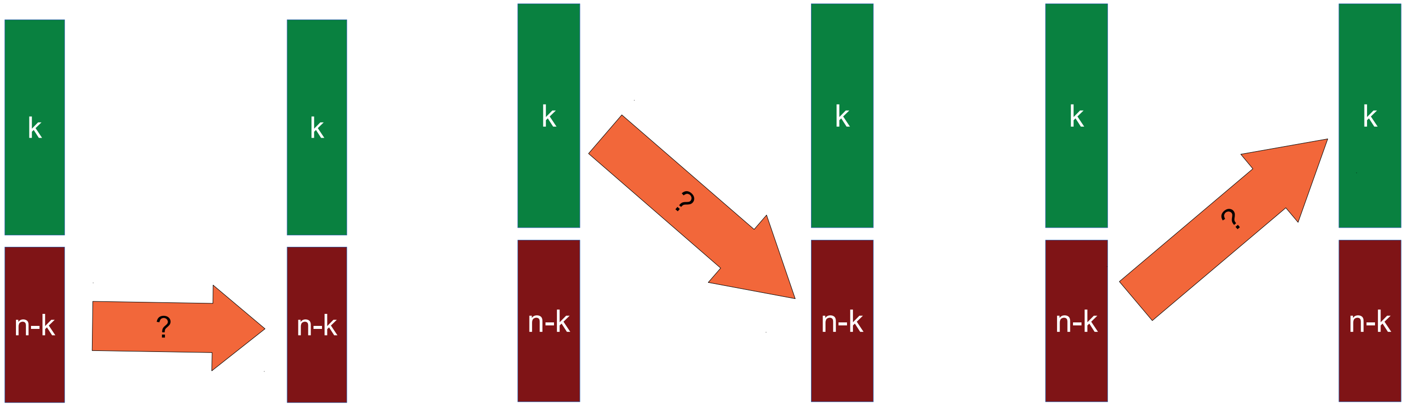
$$\forall xy(Smokes(x) \wedge Friends(x, y) \rightarrow Smokes(y))$$ n people

- If we know precisely who smokes, and there are k smokers:

Database:

$Smokes(Alice) = 1$
$Smokes(Bob) = 0$
$Smokes(Charlie) = 0$
$Smokes(Dave) = 1$
$Smokes(Eve) = 0$
$\dots$





- If we know precisely who smokes, and there are k smokers
  $\longrightarrow 2^{n^2-k(n-k)}$ models

- If we know that there are k smokers $\longrightarrow$
  $\binom{n}{k} 2^{n^2-k(n-k)}$ models

- In total $\longrightarrow$ $\sum_{k=0}^{n} \binom{n}{k} 2^{n^2-k(n-k)}$ models

Some Conclusions

- The conception and use of probability and statistical methods have direct and practical relevance in AI and KR, in particular

  There are models and methods that start from certain specific positions in this regard

- The schools of logic and probability in AI followed parallel paths for a long time

  Now we see a convergence/combination of the approaches in KR

- New kinds of "KBs" are being developed/used: MLN KBs

  There are also the new and related area of "probabilistic logic programming" [6, 4], among others

  Established connections with statistical relational learning [7, 8]

- Apart from inference, there is the problem of learning (the parameters of) those KBs

- New methods have been developed for the integration of FO logical and probabilistic reasoning, e.g. lifted inference

- Some of them appeal to existing and long-standing technology, e.g. DMBSs, with all their built-in and optimized techniques

  Some of them appeal to newer developments in data management, e.g. probabilistic DBs for approximate logical/probabilistic reasoning

- There are many applications we have not discussed here, some of them to ER and link mining [33, 32, 11, 28] and Semantic Web [9, 1]

- The future is full of interesting scientific challenges ...

http://www.scs.carleton.ca/∼bertossi

Email: bertossi@scs.carleton.ca

References

[1] Pedro Carvalho de Oliveira. Probabilistic Reasoning in the Semantic Web using Markov Logic. MSc Thesis, University of Coimbra, 2009.

[2] Yang Chen and Daisy Zhe Wang. Knowledge Expansion over Probabilistic Knowledge Bases. Proc. SIGMOD, 2014, pp. 649-660.

[3] Timothy Childers. *Philosophy & Probability*. Oxford U.P., 2013.

[4] DARPA Program on Probabilistic Programming for Advancing Machine Learning (PPAML), 2013. `https://www.fbo.gov/index?s=opportunity&mode=form&id=62b7512e2df3bc1fe220197e70a972d7&tab=core&_cview=1`

[5] Adnan Darwiche. Bayesian Networks. *Communications of the ACM*, 2010, 53(12):80-90.

[6] Luc De Raedt and Angelika Kimmig. Probabilistic Programming Concepts. In arXiv:1312.4328 [cs.PL], 2013.

[7] Thomas Dietterich et al. Structured Machine Learning: The Next Ten Years. *Machine Learning*, 2008 73:323.

[8] Pedro Domingos and Matthew Richardson. Markov Logic: A Unifying Framework for Statistical Relational Learning. In L. Getoor and B. Taskar (eds.), *Introduction to Statistical Relational Learning*. MIT Press, 2007, pp. 339-371.

[9] Pedro Domingos et al. Just Add Weights: Markov Logic for the Semantic Web. *Uncertainty Reasoning for the Semantic Web I*, Springer LNCS 5327, 2008, pp 1-25.

[10] Pedro Domingos and Daniel Lowd. Markov Logic: An Interface Layer for Artificial Intelligence. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Morgan & Claypool Publishers, 2009.

[11] Pedro Domingos et al. Markov Logic: A Language and Algorithms for Link Mining. In *Link Mining: Models, Algorithms, and Applications*, Philip S. Yu et al. (eds.), Springer, 2010, pp. 135-161.

[12] Pedro Domingos. A Few Useful Things to Know about Machine Learning. *Communications of the ACM*, 2012, 55(10):78-87.

[13] Wolfgang Gatterbauer and Dan Suciu. Approximate Lifted Inference with Probabilistic Databases. *Proc. VLDB End.*, 2015, 8(5):629-640.

[14] Eric Gribkoff, Guy Van den Broeck and Dan Suciu. Understanding the Complexity of Lifted Inference and Asymmetric Weighted Model Counting. CoRR abs/1405.3250 (2014)

[15] Eric Gribkoff, Dan Suciu and Guy Van den Broeck. Lifted Probabilistic Inference: A Guide for the Database Researcher. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2014, pp. 6-17.

[16] Joseph Y. Halpern. *Reasoning about Uncertainty*. The MIT Press, 2003.

[17] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge U.P., 2003.

[18] Gabriele Kern-Isberner, Christoph Beierle, Marc Finthammer and Matthias Thimm. Probabilistic Logics in Expert Systems: Approaches, Implementations, and Applications. Proc. DEXA'11, Springer LNCS 6860, 2011, pp. 27-46.

[19] Kristian Kersting. Lifted Probabilistic Inference. Proc. ECAI 2012, pp. 33-38.

[20] Stanley Kok and Pedro Domingos. Learning Markov Logic Network Structure via Hypergraph Lifting. Proc. ICML, 2009, pp. 505-512.

[21] Mihalkova, L. and R. J. Mooney. 2007. Bottom-up learning of Markov logic network structure. In Proc. ICML, 2007, pp. 625-632.

[22] Mathias Niepert and Guy Van den Broeck. Tractability through Exchangeability: A New Perspective on Efficient Probabilistic Inference. Proc. AAAI, 2014, pp. 2467-2475.

[23] Feng Niu, Christopher Ré, AnHai Doan, Jude Shavlik. Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS. Proc. VLDB, 2011.

[24] Jeff B. Paris. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge U.P., 1994.

[25] Jeff B. Paris. Common Sense and Maximum Entropy. *Synthese*, 1999, 117, pp. 75-93.

[26] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, revised 2nd printing, 2009.

[27] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge U.P., 2nd ed., 2009.

[28] Vibhor Rastogi, Nilesh Dalvi and Minos Garofalakis. Large-Scale Collective Entity Matching. Proc. VLDB, 2011, pp. 208-218.

[29] Raymond Reiter. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 1987, 32(1):57-95.

[30] Matthew Richardson and Pedro Domingos. Markov Logic Networks. *Machine Learning*, 2006, 62(1-2):107-136.

[31] Singla, P. and P. Domingos. 2005. Discriminative Training of Markov Logic Networks. Proc. AAAI, 2005.

[32] Parag Singla and Pedro Domingos. Object Identification with Attribute-Mediated Dependences. Proc. PKDD, 2005, Springer LNAI 3721, pp. 297308.

[33] Parag Singla and Pedro Domingos. Entity Resolution with Markov Logic. Proc. ICDM, 2006, pp. 572-582.

[34] Guy Van den Broeck. *Lifted Inference and Learning in Statistical Relational Models*. PhD Thesis, KU Leuven, 2013.

[35] Guy Van den Broeck, Wannes Meert and Adnan Darwiche. Skolemization for Weighted First-Order Model Counting. Proc. KR 2014.

[36] Frank Van Harmelen, Vladimir Lifschitz and Bruce Porter (eds.). *Handbook of Knowledge Representation*. Elsevier Science, 2008.