Markov Logic Networks for Knowledge Base Completion: A Theoretical Analysis Under the MCAR Assumption

Ondřej Kuželka
Department of CS
CTU in Prague
Prague, Czech Republic

Jesse Davis
Department of CS
KU Leuven
Leuven, Belgium

Abstract

We study the following question. We are given a knowledge base in which some facts are missing. We learn the weights of a Markov logic network using maximum likelihood estimation on this knowledge base and then use the learned Markov logic network to predict the missing facts. Assuming that the facts are missing independently and with the same probability, can we say that this approach is consistent in some precise sense? This is a non-trivial question because we are learning from only one training example. In this paper we show that the answer to this question is positive.

1 INTRODUCTION

Automatically constructed knowledge bases (KBs) such as YAGO [22] and NELL [4] contain facts, in the form of relational tuples, that have been exacted from the Web. However, these KBs are incomplete, that is, there are many facts that should be included in the KB but are not. Hence, a popular task is to attempt to automatically complete these KBs. This entails inferring which other relationships hold between the entities that appear in the KB [6, 11, 13, 20, 28, 32, 33, 34, 36]. Because the task typically only considers entities that already appear in the KB, it can be viewed as a transductive learning problem.

Given the relational nature of KBs, a popular approach to knowledge base completion is to approach it from the perspective of (probabilistic) relational learning (e.g., [20, 10, 28, 36]). A key challenge for these approaches is that the KB contains only positive examples, that is, the facts already in the KB. All other tuples should be considered as missing: if a tuple is not included in the KB, we do not know if it is true (i.e., it should be added to the KB) or false (i.e., it should be excluded from the KB). Approaches cope with this complication in various ways such as trying to learn by only considering the facts in the KB (i.e., learning from positive only data) [28], automatically inferring negative examples [4, 10] or explicitly reasoning about the missing data [36]. A drawback to these approaches is that they are ad-hoc, and lack theoretical justification or guarantees.

In this paper we theoretically study the suitability of learning the weights of a Markov logic network from a KB in the presence of missing data. After learning the weights, such an MLN could be used to infer additional facts to include in the KB. This is a challenging problem because our analysis must account for the fact that our sample only consists of a single training example: the KB. In contrast, most learning theory results assume access to multiple training examples. Our analysis focuses on the case where the available KB conforms to the missing completely at random assumption and we tackle the problem from a relational marginal point of view [18]. We show two main results. First, we show via a concentration inequality that it is possible to learn the weights of an MLN from a single example that faithfully models the unknown relational marginals, given large enough example. Second, we bound the expected difference in normalized log likelihood between the learned MLN and the optimal one.

2 PRELIMINARIES

In this section, we provide the necessary background.

2.1 FIRST ORDER LOGIC

We consider a standard function-free first-order language defined by a set of constants Δ, a set of variables $\mathcal{V}$ and for each $k \in \mathbb{N}$ a set $\mathcal{R}_k$ of k-ary predicates. To avoid confusion, variables start with lowercase letters and constants start with uppercase letters. An atom is of the form $r(a_1, ..., a_k)$ with $a_1, ..., a_k \in \Delta \cup \mathcal{V}$ and $r \in \mathcal{R}_k$. A literal is an atom or its negation. A clause is a disjunction over a finite set of literals. We assume that the variables in a clause are all universally quantified. A clause in which none of the literals contains any variables is called *ground*. The set of grounding substitutions of a clause α w.r.t. a set of constants Δ is the set $\Theta(\alpha, \Delta) = \{\vartheta_1, ..., \vartheta_m\}$ that contains substitutions to all variables occurring in α using constants from Δ. A possible world ω is represented as a set of ground atoms that are true in ω. We sometimes treat possible worlds as sets and use set-theoretic notation such as $|\omega|$ denoting the number of atoms in ω. The satisfaction relation $\models$ is defined in the usual way: $\omega \models \alpha$ means that the formula α is true in ω.

2.2 MARKOV LOGIC NETWORKS

A Markov logic network [27] (MLN) is a set of weighted first-order logic formulas (α, w), where $w \in \mathbb{R}$ and α is a function-free and quantifier-free first-order formula. The semantics are defined w.r.t. the groundings of the first-order formulas, relative to some finite set of constants Δ, called the domain. An MLN is classically seen as a template that defines a Markov random field. Specifically, an MLN Φ induces the following probability distribution on the set of possible worlds $\omega \in \Omega$: $p_\Phi(\omega) = \frac{1}{Z} \exp\left(\sum_{(\alpha,w)\in\Phi} w \cdot N(\alpha, \omega)\right)$, where $N(\alpha, \omega)$ is the number of groundings of α satisfied in ω, and Z is a normalization constant to ensure that p_Φ is a probability distribution. It turns out to be more convenient for our purposes to replace $N(\alpha, \omega)$ in the definition of MLNs by

$$Q(\alpha, \omega) = \frac{1}{|\Delta|^{|vars(\alpha)|}} \sum_{\vartheta\in\Theta(\alpha,\Delta)} \mathbb{1}(\omega \models \alpha\vartheta),$$

where $\Theta(\alpha, \Delta)$ is the set of all grounding substitutions of α's variables using constants from Δ and $\mathbb{1}(\omega \models \alpha\vartheta)$ is the indicator function, which is equal to 1 when $\alpha\vartheta$ is true in the possible world ω. Thus, $Q(\alpha, \omega)$ is the fraction of the groundings of α satisfied in ω. Hence, we will write the probability of a possible world $\omega \in \Omega$ as:

$$p_\Phi(\omega) = \frac{1}{Z} \exp\left(\sum_{(\alpha,w)\in\Phi} w \cdot Q(\alpha, \omega)\right).$$

2.3 RELATIONAL MARGINAL PROBLEMS

An alternative way to view a Markov logic network $\Phi = \{(\alpha_1, w_1), \dots, (\alpha_m, w_m)\}$ is to think of it as a maximum entropy distribution satisfying given marginal constraints $\mathbb{E}[Q(\alpha_i, .)] = \theta_i$, where $\theta_i \in [0; 1]$. Assuming we have the expected values of the formula statistics (which we might have, for instance, estimated from training data), we can define the following maximum entropy problem [18].

Relational Marginal Problem (Formulation):

$$\min_{\{P_\omega \,:\, \omega\in\Omega\}} \sum_{\omega\in\Omega} P_\omega \log P_\omega \quad \textit{s.t.} \tag{1}$$

$$\forall i = 1, \dots, m : \sum_{\omega\in\Omega} P_\omega \cdot Q(\alpha_i, \omega) = \theta_i \tag{2}$$

$$\forall \omega \in \Omega : P_\omega \geq 0, \sum_{\omega\in\Omega} P_\Omega = 1 \tag{3}$$

Here, the P_ωs are the problem's decision variables, each of which represents the probability of one possible world $\omega \in \Omega$. Line (1) is the maximum entropy criterion, which is shown here as the minimization of the negative entropy; Line (2) shows the constraints given by the statistics; and Line (3) provides the normalization constraints for the probability distribution.

Assuming there exists a feasible solution satisfying $\forall\omega : P_\omega > 0$ (we call such a solution *positive*), the optimal solution of the above maximum entropy problem is an MLN

$$P_\omega = \frac{1}{Z} \exp\left(\sum_{(\alpha_i,w_i)\in\Phi} w_i \cdot Q(\alpha_i, \omega)\right) \tag{4}$$

where the parameters $\mathbf{w} = (w_1, \dots, w_m)$ are obtained by maximizing the dual criterion

$$L(\lambda) = \sum_{\alpha_i} w_i\theta_i - \log \sum_{\omega\in\Omega} e^{\sum_{\alpha_i} w_i Q(\alpha_i,\omega)}. \tag{5}$$

This dual criterion also happens to be equivalent to the log-likelihood of the MLN (4) w.r.t. a (possibly fictitious) training example $\widehat{\omega}$ that is over the same domain Δ and satisfies $Q(\alpha_i, \widehat{\omega}) = \theta_i$ for all the formula statistics.

2.4 RELATIONAL MARGINAL POLYTOPES

Next we define relational marginal polytopes [18]. These represent the expected values for the vectors of statistics of the given formulas that are possible.

Definition 1 (Relational marginal polytope)**.** *Let Ω be a set of possible worlds and $\Phi = (\alpha_1, \dots, \alpha_m)$ be a list of formulas. We define the relational marginal polytope RMP(Φ, Ω) w.r.t. Φ as RMP$(\Phi, \Omega) = \{(x_1, \dots, x_m) \in R^l : \exists$ prob. distr. on Ω s.t. $\mathbb{E}[Q(\alpha_1, \omega)] = x_1 \wedge \dots \wedge \mathbb{E}[Q(\alpha_l, \omega)] = x_m\}$.*

Example 1. *Consider the formulas $\alpha =$ friends(x_1, x_2) and $\beta =$ friends$(x_1, x_2) \wedge$ friends$(x_2, x_3) \wedge$ friends(x_3, x_1). Let $\Delta = \{C_1, \dots, C_{100}\}$ be the set of domain elements and Ω be the respective set of possible worlds over the first-order language given by the predicate friends$/2$ and the constants from Δ. The possible worlds $\omega \in \Omega$ may be thought of as representing* *social networks. Then $Q(\alpha, \omega)$ corresponds to the "frequency" of friendships in the network and $Q(\beta, \omega)$ to the "frequency" of friendship-triangles. We can then see easily why there is, for instance, no distribution with $\mathbb{E}[Q(\alpha, \omega)] = 0$ and $\mathbb{E}[Q(\beta, \omega)] = 0.5$ (as graphs without edges cannot have a positive number of triangles). Hence, the point $(0, 1)$ will not be contained in the relational marginal polytope.*

The relational marginal polytope w.r.t. a given list of formulas $(\alpha_1, \ldots, \alpha_m)$ can be also defined as the convex hull of the set $\{(Q(\alpha_1, \omega), \ldots, Q(\alpha_m, \omega)) : \omega \in \Omega\}$.

Next we define what it means for a point to be in the η-interior of a polytope.

Definition 2 (Interiority). *Let $\eta > 0$, $\mathbf{P}$ be a polytope and $A^{=}\mathbf{x} = \mathbf{c}$ be the maximal linearly independent system of linear equations that hold for the vertices of $\mathbf{P}$ (i.e. $A^{=}$ and b define the affine subspace in which $\mathbf{P}$ "lives"). A point θ is said to be in the η-interior of $\mathbf{P}$ if $\{\theta' | A^{=}\theta' = \mathbf{c}, \|\theta' - \theta\| \leq \eta\} \subseteq \mathbf{P}$.*

We need to consider the system of linear equations $A^{=}\mathbf{x} = \mathbf{c}$ in the definition of interiority because the polytope may live in a lower dimensional subset of the given space. Our definition of interiority is also often called *relative interiority* in the literature.

If the vector of formula statistics' estimates θ is in the η-interior of the respective relational marginal polytope for some $\eta > 0$, then there always exists a positive distribution satisfying the marginal constraints given by the statistics.

We will use the following theorem from [16] which links the magnitudes of an MLN's weights and the interiority of the respective marginals.[1]

Theorem 2 (Theorem 16 in [16]). *Let Φ be a set of quantifier-free first-order logic formulas, let Ω be a set of possible worlds and $A^{=}\mathbf{x} = \mathbf{c}$ be a maximal system of linearly independent equations satisfied by the vertices of the relational marginal polytope $\mathbf{P}_R = RMP(\Phi, \Omega)$. Let θ be a point in the η-interior of $\mathbf{P}_R$. Then there is an optimal solution $\mathbf{w}^* = (w_1^*, \ldots, w_m^*)$ of the relational marginal problem constrained by the parameters θ (which is dual to the maximum-likelihood problem) such that $A^{=}\mathbf{w}^* = 0$ and any such solution satisfies $\|\mathbf{w}^*\| \leq \log|\Omega|/\eta$.*

3 LEARNING SETTING

This paper addresses the following transductive learning problem:

Given: A fixed set of constants Δ, a fixed set of relations $\mathcal{R}$ and a single sample $\widehat{\omega}$ of a knowledge base ω^* selected according to the following *data generation process*:

$$P(\omega) = \begin{cases} (1-\delta)^{|\omega|} \cdot \delta^{|\omega^*| - |\omega|} & \omega \subseteq \omega^* \\ 0 & \omega \not\subseteq \omega^* \end{cases} \tag{6}$$

where $\delta \in [0; 1]$ is the *subsampling rate*.

Do: Reconstruct ω^* from $\widehat{\omega}$.

Intuitively, this process removes from ω^* ("forgets") each of the true facts independently with probability δ. This corresponds to a relational variant of the missing data setting known as "missing completely at random" or MCAR [21]. What makes our setting more complicated is the fact that we only have one training example. Note this contrasts with the classical work on relational learning that operates on data sets consisting of many small examples, each of which can be considered to be a small relational database. For example, data about molecules fits this description: each molecule is an individual example (with a variable number of atoms and bonds) and each one can be represented using a graph. This type of data has also been considered in SRL (e.g., [14]).

If we had access to set of samples $\{\omega_1, \ldots, \omega_n\}$, where each ω_i is generated according to Equation 6, there are several approaches for reconstructing ω^*. The most obvious one would be to simply take the union of all ω_i. Another approach would be to model the probability distribution of the data generation process which has a very simple form. That is, we could model a distribution over the (independent) random variables, where we have one random variable for each possible ground atom. If we could learn an accurate enough model, we could reconstruct ω^* by taking all the ground atoms that have probability greater than certain threshold. Given a sufficient number of training examples, this would simply entail computing the empirical frequencies of the observed ground atoms.

However, since we only have one training example $\widehat{\omega}$, we need to estimate the distribution in a smarter way. The key insight underlying most of statistical relational learning is that this can be done by identifying and exploiting the structural regularities that hold in $\widehat{\omega}$. One strategy is to learn a model that maximizes the log-likelihood of the data and thereby learn the distribution. Because we are considering relational data, we can use Markov logic networks as our model class. The following example illustrates the intuition behind our approach.

Example 3. *Let the complete state of the data be $\omega^* = \{Smokes(Alice), Friends(Alice, Bob), Friends(Bob, Alice)\}$ and suppose we know that*

[1] This theorem is just a relational version of a theorem from [31].

the friendship relation is symmetric. Given the following possible world sampled by the data generation process $\widehat{\omega} = \{Friends(Alice, Bob)\}$, we could exploit our knowledge about the symmetric nature of friendship to yield the reconstruction $\omega' = \{Friends(Alice, Bob), Friends(Bob, Alice)\}$. This is obviously not perfect because it misses the $Smokes(Alice)$, but represents a step forward.

In this paper, we study from a theoretical perspective whether Markov logic networks are a suitable representation for this task. Specifically, we explore what theoretical guarantees are possible when learning the weights of an MLN from a single training example generated according to Equation 6.

4 NORMALIZED LOG-LIKELIHOOD

Ideally, we would use the KL-divergence $D_{KL}(P_{DG}||P_\Phi)$ to assess how close a distribution P_Φ modeled by MLN Φ is to the true distribution P_{DG}. However, it is more convenient to work with the log-likelihood of P_Φ, which only differs from the KL-divergence by the entropy of P_{DG}, i.e., $D_{KL}(P_{DG}||P_\Phi) = H(P_{DG}) - \mathbb{E}[L(\Phi|.)]$. Hence, we will focus on log-likelihood and note that the MLN which maximizes the expected log-likelihood is also the one that minimizes the KL-divergence.

We will need to show that the log-likelihood estimated on the one available sample converges to the expected log-likelihood in some sense. However, this poses a problem as the log-likelihood typically decreases as the size of the test example increases. This occurs even for the best possible model, which the following example illustrates.

Example 4. *Let us suppose that the true distribution P_{DG} is generated by tossing biased coins from domain $\Delta = \{Coin_1, \ldots, Coin_n\}$ that land on heads with probability 0.9. Such a distribution can clearly be modelled by an MLN Φ containing just one formula $\alpha = heads(x)$. Even if the MLN Φ modelled the distribution perfectly, the expected log-likelihood would decrease as the domain size grows bigger. Indeed, we would have $\mathbb{E}[L(\Phi|.)] = |\Delta| \cdot (0.9 \log 0.9 + 0.1 \log 0.1)$.*

To cope with this issue, we will work with the normalized log-likelihood, which is commonly done in the statistical relational learning literature (e.g. [35, 29]). We define normalized log-likelihood as follows.

Definition 3 (Normalized log-likelihood)**.** *Let Φ be an MLN, Δ be the domain and $L(\Phi, \omega)$ be the log-likelihood of Φ given ω. Let M be the number of all possible ground atoms constructed using the constants from Δ and relations from the given first-order language. Then we define the normalized log-likelihood as*

$$NL(\Phi|\omega) = \frac{1}{M} \cdot L(\Phi|\omega).$$

As each ground atom is a Boolean random variable in an MLN, this can be interpreted as the average log-likelihood per random variable.

When defining a normalized log-likelihood, it is important not to normalize by a factor that grows too quickly. If the normalization factor grew too quickly, the normalized likelihood would always converge to zero, which would make having additive bounds on the estimation error of the log-likelihood pointless. It is easy to see that our definition does not suffer from this problem because the uniform distribution over elements of Ω always has a constant normalized log-likelihood. In particular, we have $NL(\Phi|\omega) = \frac{1}{M} \log 2^{-M} = -\log 2$. Moreover, any MLN, regardless of which formulas it contains, can represent the uniform distribution because setting all the formulas' weights to 0 yields a uniform distribution. It follows that it is statistically meaningful and non-trivial to show that the normalized log-likelihood of a learned MLN will be close (in expectation) to the normalized log-likelihood of the best possible such MLN.

5 CONSISTENCY OF MLN LEARNING

Our goal is to learn maximum-likelihood weights for a Markov logic network using only a single sampled training example $\widehat{\omega}$. However, this approach only makes sense if it also guaranteed to make the distribution encoded by the Markov logic model close to the data generation distribution. For that we need to study generalization guarantees of maximum-likelihood estimation in our learning setting. In particular, we want to show that optimizing the normalized log-likelihood on the training data also leads to optimizing the expected normalized log-likelihood. If multiple independently sampled training examples $\widehat{\omega}$ were available, we could use standard tools from statistical learning theory [7]. Hence, the key challenge we address in this section is replacing these classical tools with techniques that apply when we only have the single sample $\widehat{\omega}$.

5.1 MAIN TECHNICAL RESULTS

This section presents our two main technical results. The first is Theorem 5, which provides a concentration inequality on the values of the statistics $Q(\alpha, \omega)$. As discussed in Section 2.3, a Markov logic network consisting of the formulas $\{\alpha_1, \ldots, \alpha_k\}$ can also be seen as a maximum-entropy distribution given by the sufficient statistics $Q(\alpha_i, .)$. Learning the weights for an MLN using maximum-likelihood estimation on $\widehat{\omega}$ then amounts to finding the weights w_i of the formulas α_i that make the MLN Φ (approximately) satisfy $\mathbb{E}_{\omega\sim\Phi}[Q(\alpha_i,\omega)] = Q(\alpha_i,\widehat{\omega})$ [16].

Ideally, the learned MLN should faithfully model the sufficient statistics. Therefore, we want $Q(\alpha,\widehat{\omega})$ to be concentrated close to $\mathbb{E}_{\omega\sim P_{DG}}[Q(\alpha,\omega)]$, where P_{DG} is the distribution that generates the training examples $\widehat{\omega}$. Our first theoretical result gives such a guarantee in the form of a concentration inequality that probabilistically bounds the difference $|Q(\alpha,\widehat{\omega}) - \mathbb{E}_{\omega\sim P_{DG}}[Q(\alpha,\omega)]|$.

Theorem 5. *Let P_{DG} generate training examples ω over a domain Δ according to Equation 6. Next let α be a quantifier-free formula not containing any 0-arity literals,*[2] *let $\mathcal{R}_\alpha$ be the set of relations contained in α and M be the maximum arity of relations in $\mathcal{R}_\alpha$.*[3] *Then the following inequality holds for any $\varepsilon > 0$ and $|\Delta| \geq (M+1)^M$:*

$$P_{DG}[|Q(\alpha,\omega) - \mathbb{E}[Q(\alpha,.)]| \geq \varepsilon] \leq 2\cdot\exp\left(\frac{-\varepsilon^2\cdot|\Delta|}{|\mathcal{R}_\alpha|\cdot|\alpha|^2}\right)$$

In this inequality, the domain size plays the role of the effective sample size. Interestingly, this means that the bound does not, for instance, depend on the size of the set of all possible ground atoms, which would grow much more rapidly. This is in line with other concentration inequalities that were obtained in the relational learning literature, albeit under different sampling assumptions (c.f. [18, 17]).

An absolute bound on the MLN's ability to faithfully model the relational marginals is insufficient to guarantee that its normalized log likelihood is close to optimal. Theorem 6, our second main technical result, bounds the expected difference in normalized log-likelihood between the optimal MLN Φ^* and the MLN $\widehat{\Phi}(\omega)$ learned using maximum-likelihood on a single example ω sampled according to Equation 6.

Theorem 6. *Let P_{DG} generate training examples ω over a domain Δ, $|\Delta| \geq 2$, according to Equation 6. Let us have a set of quantifier-free formulas $\{\alpha_1,\ldots,\alpha_m\}$ not containing any 0-arity literals. Let $\mathcal{H}$ ("hypothesis class") be the set of MLNs of the form $\Phi = \{(\alpha_1,w_1),\ldots,(\alpha_m,w_m)\}$ that satisfy the condition that their marginal statistics $\mathbb{E}[Q(\Phi,.)]$ are contained in the η-interior of their relational marginal polytope, where $\eta > 0$ is fixed. Then the following holds:*

$$\mathbb{E}\left[\sup_{\Phi\in\mathcal{H}}|NL(\Phi|\omega) - \mathbb{E}[NL(\Phi,.)]|\right] \leq O\left(\sqrt{\frac{\log|\Delta|}{\eta|\Delta|}}\right)$$

Denoting $\widehat{\Phi}(\omega) := \arg\max_{\Phi\in\mathcal{H}} NL(\Phi,\omega)$ (i.e., $\widehat{\Phi}(\omega)$ is the MLN obtained by maximum-likelihood learning on ω), we also have for the expected difference in normalized log-likelihood of the best possible $\Phi^* \in \mathcal{H}$ and $\widehat{\Phi}(\omega)$:

$$\sup_{\Phi^*\in\mathcal{H}}\mathbb{E}[NL(\Phi^*|.)] - \mathbb{E}_\omega\left[\mathbb{E}_{\omega'}[NL(\widehat{\Phi}(\omega),.)]\right] \leq O\left(\sqrt{\frac{\log|\Delta|}{\eta|\Delta|}}\right).$$

There are several interesting things to notice about the results from Theorem 6. First, it exhibits the same rate of convergence $O(\sqrt{\log n/n})$ that appears in VC-dimension-based bounds for the classical setting of learning from i.i.d. data [7]. The main difference is that our results depend on the size of the domain Δ as opposed to the number of training examples. This shows that the approach to knowledge base completion based on learning an MLN on the single available learning example is, in a precise sense, sound. Second, the bound is inversely proportional to the interiority parameter of the MLNs in the considered hypothesis class. This is interesting because the runtime of maximum-likelihood learning for MLNs depends polynomially on $1/\eta$ [16]. Thus, allowing smaller η's increases both the computational and sample complexity of the problem.

5.2 PROOF OF THEOREM 5

We now prove Theorem 5 using McDiarmid's inequality. This requires the following lemma, which bounds the difference in the $Q(\alpha,\omega)$ statistics for two possible worlds that only differ by the presence of one true ground atom.

Lemma 1. *Let Ω be a set of possible worlds over a domain Δ, and let $\omega_1,\omega_2 \in \Omega$ be possible worlds. If $|\omega_1 \ominus \omega_2| = 1$, where $\ominus$ denotes the symmetric difference operator,*[4] *then*

$$|Q(\alpha,\omega_1) - Q(\alpha,\omega_2)| \leq \frac{|\alpha|}{|\Delta|^a}$$

where a is the number of constants in the unique ground atom l in the symmetric difference $\omega_1 \ominus \omega_2$.

Proof. Let Θ_l be the set of those grounding substitutions $\vartheta \in \Theta(\alpha,\Delta)$ such that $\alpha\vartheta$ contains either

[2] It would be impossible to get a non-vacuous bound with 0-arity predicates present.

[3] In practice, the maximum arity will be small, e.g. it is 2 in knowledge graphs.

[4] Symmetric difference is defined as $\omega_1 \ominus \omega_2 = (\omega_1 \cup \omega_2) \setminus (\omega_1 \cap \omega_2)$.

l or $\neg l$. Then we have $|Q(\alpha,\omega_1) - Q(\alpha,\omega_2)| = \frac{1}{|\Delta|^{|vars(\alpha)|}} \left|\sum_{\vartheta \in \Theta_l} \left(\mathbb{1}(\omega_1 \models \alpha\vartheta) - \mathbb{1}(\omega_2 \models \alpha\vartheta)\right)\right| \leq \frac{1}{|\Delta|^{|vars(\alpha)|}} \cdot |\Theta_l|$. For a literal $a \in \alpha$, the number of substitutions $\vartheta \in \Theta$ such that $a\vartheta = l$ or $a\vartheta = \neg l$ is at most $|\Delta|^{|vars(\alpha)|-a}$. Using the union bound over the literals in α, we immediately obtain $|\Theta_l| \leq |\alpha| \cdot |\Delta|^{|vars(\alpha)|-a}$ which finishes the proof. □

We can now prove Theorem 5.

Proof of Theorem 5. We first redefine the random variable $Q(\alpha,\omega)$ as a function of independent Bernoulli random variables $B_1, \dots, B_{|\omega^*|}$ satisfying $P[B_i = 0] = \delta$, where δ is the subsampling rate from Equation 6. We suppose that there is some (arbitrary) ordering of the atoms in $\omega^* = \{a_1, \dots, a_{|\omega^*|}\}$ so that we could uniquely identify each B_i with an atom a_i in ω^*. Then we define a function $g : \{0,1\}^{|\omega^*|} \to 2^{\omega^*}$ as: $g(b_1, \dots, b_{\omega^*}) \mapsto \{a_i \in \omega^* | b_i = 1\}$. Finally we define $Q_\alpha(b_1, \dots, b_{\omega^*}) \triangleq Q(\alpha, g(b_1, \dots, b_{|\omega^*|}))$. It is easy to see that $Q(\alpha,\omega)$ and $Q_\alpha(B_1, \dots, B_{|\omega^*|})$ have the same distribution. As a consequence we have $P_{DG}[|Q(\alpha,\omega) - \mathbb{E}[Q(\alpha,.)]| \geq \varepsilon] = P[|Q_\alpha(B_1, \dots, B_{|\omega^*|}) - \mathbb{E}[Q_\alpha]| \geq \varepsilon]$. We also assume w.l.o.g. that ω^* contains only relations that also appear in α (since the rest of the relations in ω^* do not influence the values $Q(\alpha,\omega)$). We denote by $\mathcal{R}_\alpha \subseteq \mathcal{R}$ the set of relations present in α.

From McDiarmid's inequality [24] we have

$$P[|Q_\alpha(B_1, \dots, B_{|\omega^*|}) - \mathbb{E}[Q_\alpha]| \geq \varepsilon] \leq 2 \cdot \exp\left(\frac{-2\varepsilon^2}{\sum_{j=1}^{|\omega^*|} c_j^2}\right) \quad (7)$$

provided that $|Q_\alpha(B_1, \dots, B_j, \dots, B_{\omega^*}) - Q_\alpha(B_1, \dots, B'_j \dots, B_{\omega^*})| \leq c_j$ holds for every j and every value of B_j and B'_j. It follows from Lemma 1 that we can set $c_j := |\alpha| \cdot |\Delta|^{-A_j}$, where A_j is the number of unique constants in the atom a_j, in Inequality (7).

Let us split ω^* into disjoint subsets ω_1^*, ω_2^*, ..., ω_M^* where each ω_i^* contains all atoms from ω^* with i unique constants. Then we can write

$$\sum_{j=1}^{|\omega^*|} c_j^2 = \sum_{j=1}^{|\omega^*|} \left(|\alpha| \cdot |\Delta|^{-A_j}\right)^2 = |\omega_1^*| \cdot \left(\frac{|\alpha|}{|\Delta|}\right)^2 + \dots + |\omega_M^*| \cdot \left(\frac{|\alpha|}{|\Delta|^M}\right)^2. \quad (8)$$

We can also bound[5] every $|\omega_i^*|$ as $|\omega_i^*| \leq i^{M-1} \cdot |\mathcal{R}_\alpha| \cdot |\Delta|^i$. By substituting this into (8) and assuming that $|\Delta| \geq (M+1)^M$, we obtain

$$\sum_{j=1}^{|\omega^*|} c_j^2 \leq |\mathcal{R}_\alpha| \cdot |\alpha|^2 \cdot \left(\frac{1}{|\Delta|} + \frac{2^{M-1}}{|\Delta|^2} + \dots + \frac{M^{M-1}}{|\Delta|^M}\right)$$
$$= |\mathcal{R}_\alpha| \cdot |\alpha|^2 \cdot \frac{1}{|\Delta|} \cdot \left(1 + \frac{2^{M-1}}{|\Delta|} + \dots + \frac{M^{M-1}}{|\Delta|^{M-1}}\right)$$
$$\leq 2 \cdot \frac{|\mathcal{R}_\alpha| \cdot |\alpha|^2}{|\Delta|}.$$

Finally, plugging this into Inequality (7) finishes the proof. □

Using the same reasoning as in the proof of Theorem 5, we can obtain the following generalization of Theorem 5, which we prove in the appendix.[6]

Theorem 7. *Let P_{DG}, ω, Δ, M be as in Theorem 5. Next let $\Phi = (\alpha_1, \dots, \alpha_m)$ be a list of quantifier-free formulas not containing any 0-arity literals and let $\mathcal{R}_\Phi$ denote the set of relations contained in the formulas in Φ. Then the following inequality holds for any $\varepsilon > 0$ and $|\Delta| \geq (M+1)^M$ and $\mathbf{w} \in \mathbb{R}^m$:*

$$P_{DG}[|\langle \mathbf{w}, Q(\Phi,\omega)\rangle - \mathbb{E}[\langle \mathbf{w}, Q(\Phi,.)\rangle]| \geq \varepsilon] \leq 2 \cdot \exp\left(\frac{-\varepsilon^2 \cdot |\Delta|}{|\mathcal{R}_\Phi| \cdot \|\mathbf{w}\|^2 \cdot \left(\sum_{j=1}^m |\alpha_j|\right)^2}\right)$$

5.3 PROOF OF THEOREM 6

We now prove Theorem 6, which requires a series of lemmas. The first one is a concentration inequality for the log-likelihood.

Lemma 2. *Let $\Phi = \{(\alpha_1, w_1), \dots, (\alpha_m, w_m)\}$ be an MLN on domain Δ, $\mathcal{R}_\Phi$ be the set of relations occurring in formulas in Φ, let M be the maximum arity of relations in $\mathcal{R}_\Phi$, and let us denote $\mathbf{w} = (w_1, \dots, w_m)$. Let ω be sampled according to Equation 6. Then the following inequality holds*

$$P[|L(\Phi|\omega) - \mathbb{E}[L(\Phi|.)]| \geq \varepsilon] \leq 2 \cdot \exp\left(\frac{-\varepsilon^2 \cdot |\Delta|}{|\mathcal{R}_\Phi| \cdot \|\mathbf{w}\|^2 \cdot \left(\sum_{j=1}^m |\alpha_j|\right)^2}\right)$$

where $L(\Phi|\omega)$ is the log-likelihood of Φ given ω.

Proof. We have

$$L(\Phi|\omega) - \mathbb{E}[L(\Phi|.)] = \langle \mathbf{w}, Q(\Phi,\omega)\rangle - \mathbb{E}[\langle \mathbf{w}, Q(\Phi,.)\rangle]$$

[5] Here the factor i^{M-1} is an upper bound on the number of all possible patterns of i constants in a literal of arity M, e.g. for $M = 3$ and $i = 2$, we have the following three possible patterns: (a,a,b), (a,b,b), (a,b,a).

[6] https://bit.ly/31TC7zx

The rest of the proof of this lemma then follows directly from Theorem 7. □

Next we use Lemma 2 to derive a bound on the expected deviation of log-likelihoods from their expected values for MLNs selected from a finite set $\mathcal{H}$.

Lemma 3. *Let $\mathcal{H}$ be a finite set of MLNs given by the same set of formulas $\alpha_1, \dots, \alpha_m$ and satisfying $\|\mathbf{w}\| \le W$, where $\mathbf{w} = (w_1, \dots, w_m)$ are the weights of the MLNs' formulas. Next let $\mathcal{R}_\Phi$ be the set of relations that occur in the formulas in Φ and M be the maximum arity of relations in $\mathcal{R}_\Phi$. Finally, let ω be a possible world on the domain Δ, $|\Delta| \ge (M+1)^M$, sampled according to the data generating distribution in Equation 6. Then the following holds*

$$\mathbb{E}\left[\sup_{\Phi\in\mathcal{H}} |L(\Phi|\omega) - \mathbb{E}[L(\Phi|.)|\right] \le \sqrt{|\mathcal{R}_\Phi|}\cdot W \cdot \sum_{j=1}^{m} |\alpha_j| \sqrt{\frac{\log\left(2e|\mathcal{H}|\right)}{|\Delta|}}.$$

Proof. For notational convenience, let us first denote $Z := \sup_{\Phi\in\mathcal{H}} |L(\Phi|\omega) - \mathbb{E}[L(\Phi|.)|$ and

$$B := \frac{|\Delta|}{|\mathcal{R}_\Phi| \cdot W^2 \cdot \left(\sum_{j=1}^{m} |\alpha_j|\right)^2}.$$

Then, using the union bound and Lemma 2, we have

$$P\left[Z \ge \varepsilon\right] \le 2|\mathcal{H}| \cdot \exp\left(-\varepsilon^2 \cdot B\right).$$

Next, we have

$$\mathbb{E}[Z] \le \sqrt{\mathbb{E}[Z^2]} = \sqrt{\int_0^\infty P[Z^2 \ge t]dt}$$
$$= \sqrt{\int_0^\infty P[Z \ge \sqrt{t}]dt} \le \sqrt{u + \int_u^\infty P[Z \ge \sqrt{t}]dt}$$
$$\le \sqrt{u + \int_u^\infty 2|\mathcal{H}| \cdot \exp\left(-t \cdot B\right)dt}$$
$$= \sqrt{u + \frac{2|\mathcal{H}|}{B} \cdot \exp\left(-u \cdot B\right)}$$

The above expression is minimized for $u := \log\left(2|\mathcal{H}|\right)/B$ (note that u may be arbitrary), yielding the bound:

$$\sqrt{\frac{\log\left(2e|\mathcal{H}|\right)}{B}} = \sqrt{|\mathcal{R}_\Phi|} \cdot W \cdot \sum_{j=1}^{m} |\alpha_j| \sqrt{\frac{\log\left(2e|\mathcal{H}|\right)}{|\Delta|}}$$

□

We will use the following well-known results about the number of balls of a specific radius that are needed to cover a given subset of a metric space.

Lemma 4 (Covering number, e.g., [30])**.** *Let $\mathcal{S}$ be a subset of $\mathbb{R}^d$ of diameter at most k (i.e. for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$, $\|\mathbf{x}_1 - \mathbf{x}_2\| \le k$). Then $\mathcal{S}$ can be covered by*

$$\left(\frac{2 \cdot k \cdot \sqrt{d}}{r}\right)^d$$

balls of radius r.

Next we upper-bound the maximum change in log-likelihood when we move the weights w.

Lemma 5. *Let $\omega \in \Omega$ be a possible world on the domain Δ. Let $\Phi = \{(\alpha_1, w_1), \dots, (\alpha_m, w_m)\}$ and $\Phi' = \{(\alpha_1, w_1'), \dots, (\alpha_m, w_m')\}$ be MLNs. Then*

$$|L(\Phi|\omega) - L(\Phi'|\omega)| \le 2 \cdot \|\mathbf{w} - \mathbf{w}'\| \cdot \sqrt{|\Phi|}$$

where $\mathbf{w} = (w_1, \dots, w_m)$ and $\mathbf{w}' = (w_1', \dots, w_m')$ and $L(\Phi|\omega)$ is the log-likelihood of Φ given ω.

Proof. We have $|L(\Phi|\omega) - L(\Phi'|\omega)| =$

$$= \Bigg| \langle \mathbf{w}, Q(\Phi,\omega)\rangle - \log \sum_{\omega'\in\Omega} \exp\left(\langle \mathbf{w}, Q(\Phi,\omega')\rangle\right)$$
$$-\langle \mathbf{w}', Q(\Phi,\omega)\rangle + \log \sum_{\omega'\in\Omega} \exp\left(\langle \mathbf{w}', Q(\Phi,\omega')\rangle\right) \Bigg|$$

(using the triangle inequality)

$$\le \left|\langle \mathbf{w} - \mathbf{w}', Q(\Phi,\omega)\rangle\right|$$
$$+ \left| \log \frac{\sum_{\omega'\in\Omega} \exp\left(\langle \mathbf{w} - \mathbf{w}' + \mathbf{w}', Q(\Phi,\omega')\rangle\right)}{\sum_{\omega'\in\Omega} \exp\left(\langle \mathbf{w}', Q(\Phi,\omega')\rangle\right)} \right|$$

(using $\|Q(\Phi,\omega)\| \le \sqrt{|\Phi|}$)

$$\le \|\mathbf{w} - \mathbf{w}'\| \cdot \sqrt{|\Phi|}$$
$$+ \left| \log \frac{\sum_{\omega'\in\Omega} \exp\left(\|\mathbf{w} - \mathbf{w}'\|\sqrt{|\Phi|} + \langle \mathbf{w}', Q(\Phi,\omega')\rangle\right)}{\sum_{\omega'\in\Omega} \exp\left(\langle \mathbf{w}', Q(\Phi,\omega')\rangle\right)} \right|$$
$$= 2 \cdot \|\mathbf{w} - \mathbf{w}'\| \cdot \sqrt{|\Phi|}.$$

□

Now we finally have all the necessary machinery to prove Theorem 6. The proof relies on a covering-number based approach [30].

Proof of Theorem 6. Let us denote $d := |\Phi|$ and let $\mathcal{W} = \{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\| \le \log|\Omega|/\eta\}$ be a ball of radius $\log|\Omega|/\eta$ centered at $\mathbf{0}$; it follows from Theorem 2 that at least one optimal solution of the maximum-likelihood problem must be contained in $\mathcal{W}$. Let $\mathcal{B}$ be a finite set of points such that if we place balls of radius r in all these points then the balls cover the set $\mathcal{W}$. Let us further assume that

$$|\mathcal{B}| \leq \left(2 \cdot 2 \cdot \log|\Omega|/\eta \cdot \sqrt{d}/r\right)^d,$$

which Lemma 4 guarantees is possible. Next we assume that we are only searching for a maximum-likelihood solution among the weight vectors from $\mathcal{B}$ so that we could use Lemma 3 which expects a finite hypothesis set $\mathcal{H}$ (i.e., for us, $\mathcal{H}_\mathcal{B} := \{\{(\alpha_1, w_1), \ldots, (\alpha_m, w_m)\} | (w_1, \ldots, w_m) \in \mathcal{B}\}$). Then from Lemma 3 we have:

$$\begin{aligned}\mathbb{E}&\left[\sup_{\Phi \in \mathcal{H}_\mathcal{B}} |L(\Phi|\omega) - \mathbb{E}[L(\Phi|.)|\right] \\ &\leq \sqrt{|\mathcal{R}_\Phi|} \cdot W \cdot \sum_{j=1}^{m} |\alpha_j| \sqrt{\frac{\log(2e|\mathcal{B}|)}{|\Delta|}}. \qquad (9)\end{aligned}$$

We define $L := \sum_{j=1}^m |\alpha_j|$, which is the sum of the lengths of the formulas in Φ. Using Theorem 2 and the fact that $\log|\Omega| \leq |\mathcal{R}_\Phi| \cdot |\Delta|^A \cdot \log 2$, where A is the maximum arity among the relations in the language, we can bound the r.h.s. by:

$$R := \frac{|R_\Phi|^{1.5}|\Delta|^A L\sqrt{d}\log 2}{\eta} \cdot \sqrt{\frac{\log\left(\frac{8e|\Delta|^A\sqrt{d}}{\eta r}\right)}{|\Delta|}} \qquad (10)$$

For $\Phi' = \{(\alpha_1, w_1'), \ldots, (\alpha_m, w_m')\}$, let $\mathcal{H}_\mathcal{W}(\Phi')$ be the set of MLNs $\Phi = \{(\alpha_1, w_1), \ldots, (\alpha_m, w_m)\}$ where $(w_1, \ldots, w_m)$ is contained in the ball of radius r centered at $(w_1', \ldots, w_m')$. Now, optimizing over the set of all possible weight vectors from $\mathcal{W}$ instead of just the vectors from the set $\mathcal{H}_\mathcal{B}$, would yield the next bound:

$$\begin{aligned}\mathbb{E}&\left[\sup_{\Phi \in \mathcal{H}_\mathcal{W}} |L(\Phi|\omega) - \mathbb{E}[L(\Phi|.)]|\right] \\ &= \mathbb{E}\Bigg[\sup_{\Phi' \in \mathcal{H}_\mathcal{B}} \sup_{\Phi \in \mathcal{H}_\mathcal{W}(\Phi')} |L(\Phi|\omega) - \mathbb{E}[L(\Phi'|.)] \\ &\qquad + \mathbb{E}[L(\Phi'|.)] - \mathbb{E}[L(\Phi|.)]||] \\ &\leq \mathbb{E}\left[\sup_{\Phi' \in \mathcal{H}_\mathcal{B}} \sup_{\Phi \in \mathcal{H}_\mathcal{W}(\Phi')} |L(\Phi|\omega) - \mathbb{E}[L(\Phi'|.)]|\right] \\ &\quad + \sup_{\Phi' \in \mathcal{H}_\mathcal{B}} \sup_{\Phi \in \mathcal{H}_\mathcal{W}(\Phi')} |\mathbb{E}[L(\Phi'|.)] - \mathbb{E}[L(\Phi|.)]| \\ &\leq \mathbb{E}\Bigg[\sup_{\Phi' \in \mathcal{H}_\mathcal{B}} \sup_{\Phi \in \mathcal{H}_\mathcal{W}(\Phi')} |L(\Phi|\omega) - L(\Phi'|\omega) \\ &\qquad + L(\Phi'|\omega) - \mathbb{E}[L(\Phi'|.)]||] \\ &\quad + \sup_{\Phi' \in \mathcal{H}_\mathcal{B}} \sup_{\Phi \in \mathcal{H}_\mathcal{W}(\Phi')} |\mathbb{E}[L(\Phi'|.)] - \mathbb{E}[L(\Phi|.)]| \\ &\leq \mathbb{E}\left[\sup_{\Phi' \in \mathcal{H}_\mathcal{B}} \sup_{\Phi \in \mathcal{H}_\mathcal{W}(\Phi')} |L(\Phi|\omega) - L(\Phi'|\omega)|\right] \\ &\quad + \mathbb{E}\left[\sup_{\Phi' \in \mathcal{H}_\mathcal{B}} |L(\Phi'|\omega) - \mathbb{E}[L(\Phi'|.)]|\right] \\ &\quad + \sup_{\Phi' \in \mathcal{H}_\mathcal{B}} \sup_{\Phi \in \mathcal{H}_\mathcal{W}(\Phi')} |\mathbb{E}[L(\Phi'|.)] - \mathbb{E}[L(\Phi|.)]|\end{aligned}$$

Next, to bound the first and third term, we use Lemma 5 and, to bound the second term, we use Equations (9) and (10). After setting

$$r := \frac{1}{\eta\sqrt{|\Delta|}}$$

and simplifying, this gives us the following bound:

$$\begin{aligned}\mathbb{E}&\left[\sup_{\Phi \in \mathcal{H}_\mathcal{W}} |L(\Phi|\omega) - \mathbb{E}[L(\Phi|.)]|\right] \\ &\leq \frac{|\mathcal{R}_\Phi|^{1.5}|\Delta|^A L \log 2\sqrt{d\log(8e|\Delta|^{A+0.5}\sqrt{d})} + 4\sqrt{d}}{\eta\sqrt{|\Delta|}} \\ &= O\left(|\Delta|^A \cdot \sqrt{\frac{\log|\Delta|}{|\Delta|}}\right).\end{aligned}$$

For the expected error of normalized log-likelihood we then obtain

$$\mathbb{E}\left[\sup_{\Phi \in \mathcal{H}_\mathcal{B}} |NL(\Phi|\omega) - \mathbb{E}[NL(\Phi|.)|\right] \leq O\left(\sqrt{\frac{\log|\Delta|}{\eta|\Delta|}}\right)$$

which finishes the proof of the first part of the theorem.

The second part of the theorem follows easily from the same argument as the first part. We use the following inequality (e.g. Lemma 8.2 in [7]) that holds for every ω:

$$\begin{aligned}\sup_{\Phi^* \in \mathcal{H}} &\mathbb{E}[L(\Phi^*|.)] - \mathbb{E}[L(\widehat{\Phi}(\omega), .)] \\ &\leq 2 \cdot \sup_{\Phi \in \mathcal{H}} |L(\Phi, \omega) - \mathbb{E}[L(\Phi, .)]|.\end{aligned}$$

□

6 RELATED WORK

This work is related to learning from missing data [21], which has not received much attention in the statistical relational learning literature. A standard approach is to perform structural EM [15] for learning both the weights and the structure in the presence of missing data. The work of Neumann et al. [26] is similar to ours in that it assumes positive-only data in a transductive setting. It attempts to infer negative examples, but is concerned with clustering and not weight learning.

The problem can also be viewed from the prism of learning from positive and unlabeled (PU) data [1] or learning from purely positive data [5, 23, 25]. In our setting, ground atoms either are in the knowledge base (i.e., are positive examples) and all other ground atoms are unlabeled, i.e., they may or may not belong in the knowledge base. One way to view our work is as a multi-target variant of the non-traditional classifiers used in positive and unlabeled learning [9] for propositional data. Recently, there has been some work on PU learning for relational data [2, 3, 10, 36]. All of these approaches are inductive and focus on learning rules (i.e., structure learning from the MLN perspective), possibly with an associated weight or confidence, from data. Some work focuses on predicting a single target predicate [2, 3] whereas other work focuses on a multi-predicate approach [10, 36]. Both [2, 36] look at trying to understand the data generation procedure, but neither of them study it from a theoretical perspective.

Several works have explored statistical learning theory for SRL. For instance, Xiang and Neville [35] studied the consistency of estimation. However, in their setting the relational graph is fixed and one only predicts the labels of vertices by exploiting the relational structure to make the predictions. Additionally, there are rather strong assumptions on the sequence of relational graphs as their size tends to infinity: (i) bounded degree and (ii) weak dependence. Our work does not require bounded degree assumptions. Note that the interiority parameter that appears in our bounds from Theorem 6 is related to weak dependence assumptions, however, it is more explicit, as it directly refers to properties of the marginal statistics. Under similar assumptions, in particular assuming a fixed relational graph, He and Zhang [12] extended the results of Xiang and Neville to the non-asymptotic setting.

Dhurandhar and Dobra [8] derived Hoeffding-type inequalities for classifiers trained from relational data. However, these inequalities, which are based on the restriction on the independent interactions of data points, cannot be applied to solve the problems tackled in this paper. In particular, these bounds also assume a fixed relational graph. Finally, recent work has proposed VC-dimension based bounds for relational learning [19]. However, that work only provides bounds for the sufficient statistics under the assumption that the training example is induced by a subset sampled from the domain uniformly. More importantly, it requires that there are no missing facts, which makes it inapplicable to our setting.

7 CONCLUSIONS

This paper studied the question of whether it makes sense to use MLNs for knowledge base completion in the most naive way: first learning their weights on the given, incomplete, knowledge base, treating it as if it were complete (i.e. using the so-called *closed-world assumption*), and then using the learned MLN for prediction on the same knowledge base to infer missing facts. For this approach to make sense, a necessary condition is that the learned distribution represented by the MLN should be as close to the data generation distribution as possible. In particular, maximizing the log-likelihood on training data should lead to maximizing the expected log-likelihood of the MLN model. Under the assumption that facts are missing from the knowledge base *completely at random*, we showed that the normalized log-likelihoods of the learned models converge to the optimal ones in expectation with the rate $O(\sqrt{\log|\Delta|/(\eta|\Delta|)})$ where Δ is the set of domain elements and η is a parameter measuring how extreme the values of the sufficient statistics of the learned MLNs may be (the smaller the value of η, the more extreme the statistics may be). We have also derived bounds on the estimated values of the sufficient statistics. It follows from our results that the naive strategy for knowledge base completion using MLNs that we considered here is, perhaps a bit surprisingly, justifiable by theoretical arguments.

Acknowledgements

OK is supported by the OP VVV project *CZ.02.1.01/0.0/0.0/16_019/0000765* “Research Center for Informatics”. JD is partially supported by KU Leuven Research Fund (C14/17/07, C32/17/036), Research Foundation - Flanders (EOS No. 30992574, G0D8819N) and VLAIO-SBO grant HYMOP (150033).

References

[1] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *arXiv preprint arXiv:1811.04820*, 2018.

[2] Jessa Bekker and Jesse Davis. Positive and unlabeled relational classification through label frequency estimation. In *Inductive Logic Programming*, pages 16–30. Springer, 2018.

[3] Hendrik Blockeel. PU-learning disjunctive concepts in ILP. In *ILP 2017 late breaking papers*, 2017.

[4] Andrew Carlson, Justin Betteridge, and Bryan Kisiel. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the 24th Con-*

ference on Artificial Intelligence (AAAI'10), pages 1306–1313, 2010.

[5] James Cussens. Using prior probabilities and density estimation for relational classification. In *Proceedings of the 8th International Conference on Inductive Logic Programming*, pages 106–115, 1998.

[6] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2D knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1811–1818, 2018.

[7] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[8] Amit Dhurandhar and Alin Dobra. Distribution-free bounds for relational classification. *Knowledge and information systems*, 31(1):55–78, 2012.

[9] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pages 213–220, 2008.

[10] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. Fast rule mining in ontological knowledge bases with AMIE+. *The VLDB JournalThe International Journal on Very Large Data Bases*, 24(6):707–730, 2015.

[11] Matt Gardner, Partha Talukdar, Jayant Krishnamurthy, and Tom Mitchell. Incorporating vector space similarity in random walk inference over knowledge bases. In *Proceedings of EMNLP*, 2014.

[12] Peng He and Changshui Zhang. Non-asymptotic analysis of relational learning with one network. In *Artificial Intelligence and Statistics*, pages 320–327, 2014.

[13] Wenqiang He, Yansong Feng, Lei Zou, and Dongyan Zhao. Knowledge base completion using matrix factorization. In *Web Technologies and Applications - 17th Asia-PacificWeb Conference (APWeb'15)*, pages 768–780, 2015.

[14] Hassan Khosravi, Oliver Schulte, Tong Man, Xiaoyuan Xu, and Bahareh Bina. Structure learning for markov logic networks with many descriptive attributes. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[15] Tushar Khot, Sriraam Natarajan, Kristian Kersting, and Jude W. Shavlik. Gradient-based boosting for statistical relational learning: the markov logic network and missing data cases. *Machine Learning*, 100(1):75–100, 2015.

[16] Ondrej Kuzelka and Vyacheslav Kungurtsev. Lifted weight learning of Markov logic networks revisited. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS-19)*, 2019.

[17] Ondrej Kuzelka, Yuyi Wang, Jesse Davis, and Steven Schockaert. PAC-reasoning in relational domains. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018*, pages 927–936, 2018.

[18] Ondrej Kuzelka, Yuyi Wang, Jesse Davis, and Steven Schockaert. Relational marginal problems: Theory and estimation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. AAAI Press, 2018.

[19] Ondrej Kuzelka, Yuyi Wang, and Steven Schockaert. VC-dimension based generalization bounds for relational learning. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2018.

[20] Ni Lao, Amarnag Subramanya, Fernando Pereira, and William W Cohen. Reading the web with learned syntactic-semantic inference rules. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1017–1026. Association for Computational Linguistics, 2012.

[21] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

[22] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek. YAGO3 : A Knowledge Base from Multilingual Wikipedias. In *Proceedings of the Conference on Innovative Data Systems Research, CIDR '15*, 2015.

[23] Eric McCreath and Arun Sharma. ILP with noise and fixed example size: A bayesian approach. In *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence*, pages 1310–1315, 1997.

[24] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[25] Stephen Muggleton. Learning from positive data. In *Selected Papers from the 6th International Workshop on Inductive Logic Programming*, pages 358–376, 1996.

[26] Marion Neumann, Babak Ahmadi, and Kristian Kersting. Markov logic sets: Towards lifted information retrieval using pagerank and label propagation. In

Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.

[27] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

[28] Stefan Schoenmackers, Oren Etzioni, Daniel S Weld, and Jesse Davis. Learning first-order horn clauses from web text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1088–1098. Association for Computational Linguistics, 2010.

[29] Oliver Schulte and Sajjad Gholami. Locally consistent Bayesian network scores for multi-relational data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 2693–2700, 2017.

[30] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[31] Mohit Singh and Nisheeth K Vishnoi. Entropy, optimization and counting. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing (STOC)*, pages 50–59. ACM, 2014.

[32] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934, 2013.

[33] Théo Trouillon, Christopher R. Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. Knowledge graph completion via complex tensor factorization. *Journal of Machine Learning Research*, 18:130:1–130:38, 2017.

[34] William Yang Wang and William W Cohen. Learning first-order logic embeddings via matrix factorization. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, pages 2132–2138, 2016.

[35] Rongjing Xiang and Jennifer Neville. Relational learning with one network: An asymptotic analysis. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 779–788, 2011.

[36] Kaja Zupanc and Jesse Davis. Estimating rule quality for knowledge base completion with the relationship between coverage assumption. In *Proceedings of the Web Conference 2018*, pages 1–9, 2018.