A Scalable Probabilistic Description Logic Approach for Semantic Link Prediction

José Eduardo Ochoa Luna[1] **, Kate Revoredo**[2] **, Fabio Gagliardi Cozman**[1]

[1] Escola Politécnica, Universidade de São Paulo,
Av. Prof. Mello Morais 2231, São Paulo - SP, Brazil

[2]Departamento de Informática Aplicada, Unirio
Av. Pasteur, 458, Rio de Janeiro, RJ, Brazil

eduardo.ol@gmail.com,katerevoredo@uniriotec.br,fgcozman@usp.br

Abstract. *Predicting potential links between unconnected nodes in a network, as collaboration networks, is a problem of great practical interest. Link prediction is mostly based on graph-based features and recently, on approaches that consider semantics of the domain. However, there is uncertainty in these predictions and considering it, can improve the prediction results. In this paper, we propose an algorithm for link prediction that uses a probabilistic ontology described with the probabilistic description logic* CR$\mathcal{ALC}$*. Moreover, our approach is scalable through a combination with graph-based features. A dataset based on the Lattes curriculum platform is used to evaluate empirically our proposal.*

1. Introduction

Many social, biological, and information systems can be well described by networks, where nodes represent objects (individuals), and links denote the relations or interactions between nodes. Predicting a possible link in a network is an interesting issue that has recently gained attention, due to the growing interest in social networks. For instance, one may be interested in finding potential friendship between two persons in a social network, or a potential collaboration between two researchers. Thus link prediction [Liben-Nowell and Kleinberg 2003, Taskar et al. 2003] aims at predicting whether two nodes (i.e., people) should be connected given that we know previous information about their relationships or interests. A common approach is to exploit the network structure, where numerical information about nodes is analyzed [Liben-Nowell and Kleinberg 2003, Taskar et al. 2003, Hsu et al. 2006]. However, knowledge about the objects represented in the nodes can improve prediction results. For instance consider that the researchers *Joe* and *Mike* do not have a publication in common, thus they do not share a link in a collaboration network. Moreover, graph features do not indicate a potential link between them. However, they have published in the same journal and they both teach the same course in their respectively universities. This information can be an indication of a potential collaboration between them. Given this, approaches that are based on the semantics related to the domain of the objects represented by the nodes [Wohlfarth and Ichise 2008, Sachan and Ichise 2011] have been proposed. In some of them, an ontology modeling the domain and the object interests were used in the prediction task.

However, there is uncertainty in such predictions. Often, it is not possible to guarantee the relationship between two objects (nodes). This is maybe due to the fact that information about the domain is incomplete. Thus, it would be interesting if link prediction approaches could handle the *probability* of a link conditioned on the information about the domain. In our example, knowing that the probability of the relationship between Joe and $Mike$ conditioned on the knowledge of them publishing in the same journal and teaching the same course is high implies a link between them in the network; otherwise, a link is not suggested. In graph-based approaches, probabilistic models learned through machine learning algorithms were used for link prediction. Some examples of probabilistic models are Probabilistic Relational Model (PRM) [Friedman et al. 1999], Probabilistic Entity Relationship Model (PERM) [Heckerman et al. 2004] and Stochastic Relational Model (SRM) [Yu et al. 2006]. On approaches based on semantics we claim that ontologies must be used to model the domain. Therefore, to model uncertainty, probabilistic approaches, such as probabilistic ontologies, must be considered.

An ontology can be represented through a description logic [Baader and Nutt 2002], which is typically a decidable fragment of first-order logic that tries to reach a practical balance between expressivity and complexity. To encode uncertainty, a probabilistic description logic (PDL) must be contemplated. The literature contains a number of proposals for PDLs [Heinsohn 1994, Jaeger 1994, Sebastiani 1994]. In this paper we adopt a recently proposed PDL, called Credal $\mathcal{ALC}$ (CR$\mathcal{ALC}$) [Cozman and Polastro 2008, Polastro and Cozman 2008, Cozman and Polastro 2009], that extends the popular logic $\mathcal{ALC}$ [Baader and Nutt 2002]. In CR$\mathcal{ALC}$ one can specify sentences such as $P(\mathsf{Professor}|\mathsf{Researcher}) = 0.4$, indicating the probability that an element of the domain is a Professor given that it is a Researcher. These sentences are called *probabilistic inclusions*. Exact and approximate inference algorithms that deal with probabilistic inclusions have been proposed [Cozman and Polastro 2008, Cozman and Polastro 2009], using ideas inherited from the theory of Relational Bayesian Networks (RBN)[Jaeger 2002].

In this paper, we propose to use a probabilistic ontology defined with the PDL CR$\mathcal{ALC}$ for semantic link prediction. Moreover, we combine it with graph-based approaches in order to make our proposal scalable.

The paper is organized as follows. Section 2 reviews basic concepts of PDLs and CR$\mathcal{ALC}$. Section 3 presents our algorithm for semantic link prediction through the PDL CR$\mathcal{ALC}$. Experiments are discussed in Section 4, and Section 5 concludes the paper.

2. Probabilistic Description Logics and CR$\mathcal{ALC}$

Description logics (DLs) form a family of representation languages that are typically decidable fragments of first order logic (FOL) [Baader and Nutt 2002]. Knowledge is expressed in terms of *individuals*, *concepts*, and *roles*. The semantics of a description is given by a *domain* $\mathcal{D}$ (a set) and an *interpretation* $\cdot^{\mathcal{I}}$ (a functor). Individuals represent objects through names from a set $N_{\mathsf{I}} = \{a, b, \ldots\}$. Each *concept* in the set $N_{\mathsf{C}} = \{C, D, \ldots\}$ is interpreted as a subset of a domain $\mathcal{D}$. Each *role* in the set $N_{\mathsf{R}} = \{r, s, \ldots\}$ is interpreted as a binary relation on the domain.

Several probabilistic descriptions logics (PDLs) have appeared in the literature. Heinsohn [Heinsohn 1994], Jaeger [Jaeger 1994] and Sebastiani [Sebastiani 1994] con-

sider probabilistic inclusion axioms such as $P_{\mathcal{D}}(\mathsf{Professor}) = \alpha$, meaning that a randomly selected object is a Professor with probability α. This characterizes a *domain-based* semantic: probabilities are assigned to subsets of the domain $\mathcal{D}$. Sebastiani also allows inclusions such as $P(\mathsf{Professor}(\mathsf{John})) = \alpha$, specifying probabilities over the interpretations themselves. For example, one interprets $P(\mathsf{Professor}(\mathsf{John})) = 0.001$ as assigning 0.001 to be the probability of the set of interpretations where John is a Professor. This characterizes an *interpretation-based* semantic.

The PDL CR$\mathcal{ALC}$ is a probabilistic extension of the DL $\mathcal{ALC}$ that adopts an interpretation-based semantic. It keeps all constructors of $\mathcal{ALC}$, but only allows concept names on the left hand side of inclusions/definitions. Additionally, in CR$\mathcal{ALC}$ one can have probabilistic inclusions such as $P(C|D) = \alpha$ or $P(r) = \beta$ for concepts C and D, and for role r. If the interpretation of D is the whole domain, then we simply write $P(C) = \alpha$. The semantic of these inclusions is roughly (a formal definition can be found in [Cozman and Polastro 2009]) given by:

$$\forall x \in \mathcal{D} \ : \ P(C(x)|D(x)) = \alpha,$$

$$\forall x \in \mathcal{D}, y \in \mathcal{D} \ : \ P(r(x,y)) = \beta.$$

We assume that every terminology is acyclic; no concept uses itself. This assumption allows one to represent any terminology $\mathcal{T}$ through a directed acyclic graph. Such a graph, denoted by $\mathcal{G}(\mathcal{T})$, has each concept name and role name as a node, and if a concept C directly uses concept D, that is if C and D appear respectively in the left and right hand sides of an inclusion/definition, then D is a *parent* of C in $\mathcal{G}(\mathcal{T})$. Each existential restriction $\exists r.C$ and value restriction $\forall r.C$ is added to the graph $\mathcal{G}(\mathcal{T})$ as nodes, with an edge from r and C to each restriction directly using it. Each restriction node is a *deterministic* node in that its value is completely determined by its parents. The graph $\mathcal{G}(\mathcal{T})$ is a Relational Bayesian Network (RBN) [Jaeger 2001].

Example 1. Consider a terminology $\mathcal{T}_1$ with concepts $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$. Suppose $P(\mathsf{A}) = 0.9, \mathsf{B} \sqsubseteq \mathsf{A}, \mathsf{C} \sqsubseteq \mathsf{B} \sqcup \exists r.\mathsf{D}, P(\mathsf{B}|\mathsf{A}) = 0.45, P(\mathsf{C}|\mathsf{B} \sqcup \exists r.\mathsf{D}) = 0.5$, and $P(\mathsf{D}|\forall r.\mathsf{A}) = 0.6$. The last three assessments specify beliefs about partial overlap among concepts. Suppose also $P(\mathsf{D}|\neg\forall r.\mathsf{A}) = \epsilon \approx 0$ (conveying the existence of exceptions to the inclusion of D in $\forall r.\mathsf{A}$). Figure 1 in the left depicts $\mathcal{G}(\mathcal{T})$, while the graph in the right illustrates the grounding of $\mathcal{G}(\mathcal{T})$ for a domain with two individuals ($\mathcal{D} = \{a, b\}$).

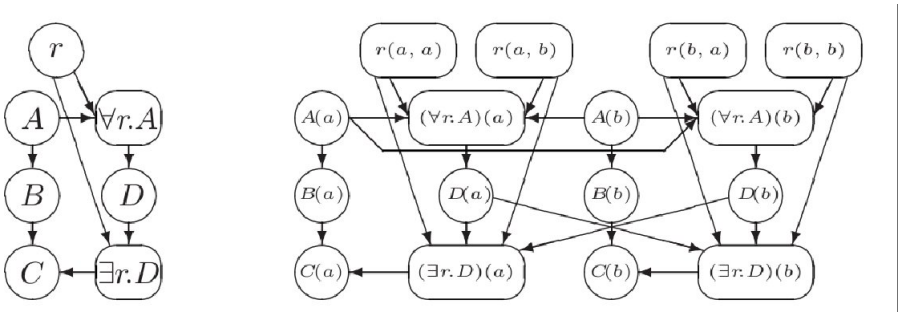

Figure 1. $\mathcal{G}(\mathcal{T})$ for terminology $\mathcal{T}$ in Example 1 and its grounding for domain$\mathcal{D} = \{a, b\}$.

The semantic of CR$\mathcal{ALC}$ is based on probability measures over the space of interpretations, for a fixed domain. Inferences, such as $P(\mathsf{A}_o(\mathsf{a}_0)|\mathcal{A})$ for an ABox $\mathcal{A}$, can be computed by propositionalization, generating a grounding RBN, where one slice is built for each individual. Therefore, not always exact probabilistic inference is possible. In [Cozman and Polastro 2009], a first order loopy propagation algorithm was proposed for approximate calculations.

3. Link Prediction with CR$\mathcal{ALC}$

In this section we describe how to apply the PDL CR$\mathcal{ALC}$ for semantic link prediction in a scalable way.

As in graph-based approaches, nodes are entities (represented by letters $a, b, c, \ldots$) in a network $\mathcal{N}$, and we are interested in defining whether a link between a and b is suitable given that there is no link between these nodes in $\mathcal{N}$. Interests, i.e., semantics between the nodes is modeled through a probabilistic ontology represented by the PDL CR$\mathcal{ALC}$. In addition, graph path information is used to improve probabilistic inference. In summary, the semantic link prediction task proposed in this paper can be described as:

Given:

- a network $\mathcal{N}$ defining relationship between objects;
- an ontology $\mathcal{O}$ in CR$\mathcal{ALC}$ describing the domain of the objects;
- the ontology concept $\mathcal{C}$ that defines the semantics of the network objects;
- the ontology role $r(\_, \_)$ that defines the semantics of the relationship between network objects;

Find:

- a revised network $\mathcal{N}_f$ with new relationship between objects.

The proposed algorithm for link prediction receives a network of a specific domain. For instance, in a co-authorship network the nodes represent researchers and the relationship can have the semantic "has a publication with" or "is advised by". Therefore, the ontology represented by CR$\mathcal{ALC}$ describes the domain of publications between researchers, having concepts like Researcher, Publication, StrongRelatedResearcher and NearCollaborator and roles like hasPublication, hasSameInstitution and sharePublication. This ontology can be learned automatically through a learning algorithm as the ones proposed in [Ochoa-Luna et al. 2010, Revoredo et al. 2010, Ochoa-Luna et al. 2011]. Thus, the nodes represent instances of one of the concepts described in the PDL CR$\mathcal{ALC}$ and the semantic of the links is described by one of the roles in the PDL CR$\mathcal{ALC}$. These concept and role must be informed as inputs to the proposed algorithm. The link prediction algorithm is described in Algorithm 1.

The algorithm starts by looking for all pairs of instances of the concept $\mathcal{C}$ defined as the concept that provides the semantic for the network nodes — this is a general setting, as a rule the set of possible pairs is restricted. For each pair, it checks whether a link between the corresponding nodes exists in the network. If not the probability of the link is calculated through the probability of the defined role conditioned on evidences (step 5). The evidences are provided by the instances of the ontology. The number of instances in an ontology has a great impact in inference. Usually one considers that more instances better inference. However, evidences for different individuals can turn out the inference

Require: a network $\mathcal{N}$, an ontology $\mathcal{O}$, the role $r(\_,\_)$ representing the semantic of the network link, the concept $\mathcal{C}$ describing the objects of the network and a $threshold$.
Ensure: a revised network $\mathcal{N}_f$
1: define $\mathcal{N}_f$ as $\mathcal{N}$;
2: **for all** pair of instances (a, b) of concept $\mathcal{C}$ **do**
3: **if** does not exist a link between nodes a and b in the network $\mathcal{N}$ **then**
4: compute $evidence$ based on a, b and nodes in their path;
5: infer probability $P(r(a, b)|evidence)$ using the RBN created through the ontology $\mathcal{O}$;
6: **if** $P(r(a, b)|evidence) > threshold$ **then**
7: add a link between a and b in network $\mathcal{N}_f$;
8: **end if**
9: **end if**
10: **end for**
Algorithm 1: Algorithm for link prediction through CR$\mathcal{ALC}$.

process computationally expensive, since in a RBN a slice is created for each individual, and then inference should be done for each slice. In [Cozman and Polastro 2009], an approximate inference algorithm was proposed where all slices without evidence are consolidated in a unique slice, thus making inference feasible in real domains. Therefore, less individuals with evidence faster inference is. From another perspective we are interested in predicting a relationship between two individuals, a and b. Therefore, evidences for these two individuals and other individuals strongly related to them are more relevant for link prediction than evidences from other individuals in the network. Thus, in this paper we propose to consider evidences about a, b and the individuals in their path, which makes the link prediction problem scalable for large networks. Therefore, in step 4 the nodes (individuals) belonging to the path between a and b are found. The inference is then performed through CR$\mathcal{ALC}$ lifted variational method on ontology $\mathcal{O}$. If the probability inferred is greater than a threshold then the corresponding link is added to the network. Alternatively, when the threshold to be considered is not known a priori, a rank of the inferred links based on their probability is done and the top-k, where k would be a parameter, are chosen.

4. Experiments

In order to evaluate our proposal empirical experiments were performed. To do so, a real world dataset was used and our algorithm was combined with state-of-the-art measures on a classification model for link prediction. This section reports on steps involved in this process.

4.1. Scenario Description

The Lattes Curriculum Platform is the public repository of Brazilian scientic curriculum which is comprised by approximately a million of registered researchers. Information is given in HTML format, and ranges from personal information such as name and address to a list of publications, examination boards' participations, main research areas, main research projects and advising information. There is implicit relational information in these

HTML pages, for instance co-authoring networks, advising/adviser links, relationships on institutions. We have randomly selected a set of 1100 researchers from engineering and math backgrounds. Based on assertional data about these researchers a probabilistic ontology has been learned. This ontology has also been extended with some probabilistic roles — learning is mainly addressed to probabilistic inclusions and concepts. The revised ontology is as follows.

	$P(\mathsf{Publication}) = 0.3$
	$P(\mathsf{Board}) = 0.33$
	$P(\mathsf{sharePublication}) = 0.22$
	$P(\mathsf{wasAdvised}) = 0.05$
	$P(\mathsf{hasSameInstitution}) = 0.14$
	$P(\mathsf{sameExaminationBoard}) = 0.31$
$\mathsf{ResearcherLattes} \equiv$	Person
	$\sqcap(\exists\mathsf{hasPublication.Publication}$
	$\sqcap\exists\mathsf{advises.Person} \sqcap \exists\mathsf{participate.Board})$
$P(\mathsf{PublicationCollaborator}$	$\mid \mathsf{Researcher} \sqcap \exists\mathsf{sharePublication.Researcher}) = 0.91$
$P(\mathsf{SupervisionCollaborator}$	$\mid \mathsf{Researcher} \sqcap \exists\mathsf{wasAdvised.Researcher}) = 0.94$
$P(\mathsf{SameInstitution}$	$\mid \mathsf{Researcher} \sqcap \exists\mathsf{hasSameInstitution.Researcher}) = 0.92$
$P(\mathsf{SameBoard}$	$\mid \mathsf{Researcher}\sqcap$
	$\exists\mathsf{sameExaminationBoard.Researcher}) = 0.95$
$P(\mathsf{NearCollaborator}$	$\mid \mathsf{Researcher} \sqcap \exists\mathsf{sharePublication.}\exists\mathsf{hasSameInstitution.}$
	$\exists\mathsf{sharePublication.Researcher}) = 0.95$
$\mathsf{FacultyNearCollaborator} \equiv$	$\mathsf{NearCollaborator}$
	$\sqcap\ \exists\mathsf{sameExaminationBoard.Researcher}$
$P(\mathsf{NullMobilityResearcher}$	$\mid \mathsf{Researcher} \sqcap \exists\mathsf{wasAdvised.}$
	$\exists\mathsf{hasSameInstitution.Researcher}) = 0.98$
$\mathsf{StrongRelatedResearcher} \equiv$	$\mathsf{Researcher}$
	$\sqcap\ (\exists\mathsf{sharePublication.Researcher} \sqcap$
	$\exists\mathsf{wasAdvised.Researcher})$
$\mathsf{InheritedResearcher} \equiv$	$\mathsf{Researcher}$
	$\sqcap\ (\exists\mathsf{sameExaminationBoard.Researcher} \sqcap$
	$\exists\mathsf{wasAdvised.Researcher})$

In this probabilistic ontology concepts and probabilistic inclusions denote mutual research interests. For instance, a PublicationCollaborator inclusion refers to Researchers who share a Publication, thus relates two nodes (Researcher) in a collaboration graph. Therefore, the concept Researcher and the role sharePublication are inputs to the algorithm we showed in Algorithm 1. To perform inferences and therefore to obtain link predictions we resort to the variational algorithm in CR$\mathcal{ALC}$.

In addition, we have also defined a collaboration network based on publication coauthoring. Topological graph information was computed accordingly. Figure 2 depicts a subset of collaborations among researchers.

If we carefully inspect this collaboration graph we could be interested, for instance, in predicting links among researchers from different groups. Since filling form is prone to errors, there is uncertainty regarding real collaborations. Thus, in Figure 2 one

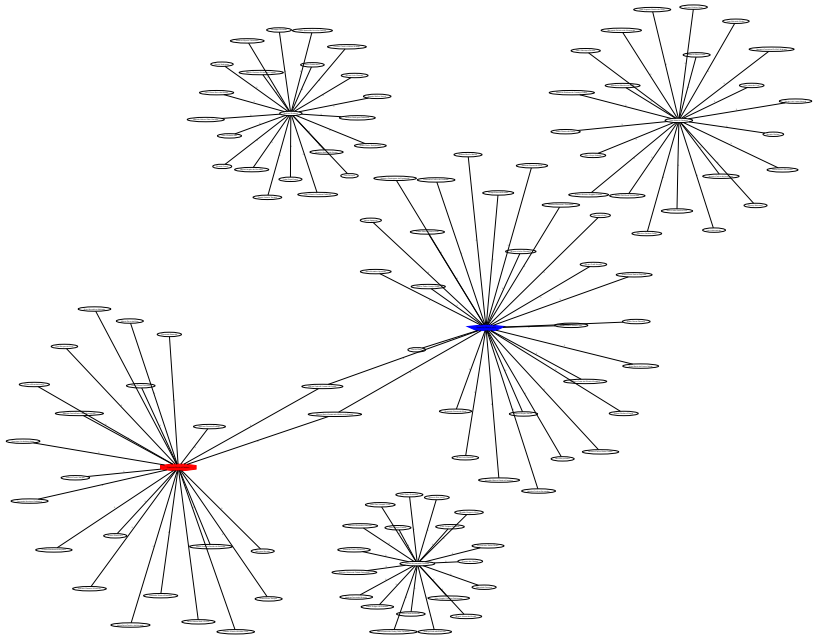

Figure 2. Lattes collaboration network.

could further investigate whether a link between researcher R (red octagon node) and the researcher B (blue polygon node) is suitable.

In order to infer this, the probability of a possible link between R and B is calculated, $P(link(R, B)|E)$, where E denotes evidence about researchers such as publications, institution, examination board participations and so on. The role sharePublication is the one defining the semantic of the links in the graph. Therefore, it is through it that we must calculate $P(link(R, B)|E)$. Since the concept PublicationCollaborator is defined by the role sharePublication and considering as evidence $\mathsf{Researcher}(R) \sqcap \exists \mathsf{hasSameInstitution}.\mathsf{Researcher}(B)$ one can infer $P(link(R, B)|E)$ through:

$$P(\mathsf{PublicationCollaborator(R)} \mid \mathsf{Researcher(R)} \sqcap \exists \mathsf{hasSameInstitution}.\mathsf{Researcher}(B)) = 0.57.$$

If we took a threshold of 0.60, the link between R and B would not be included.

One could gain more evidence, such as information about nodes that indirectly connect these two groups (Figure 2), denoted by I_1, I_2. The inference would be

$$P(\mathsf{PublicationCollaborator(R)} \mid \mathsf{Researcher(R)} \sqcap \exists \mathsf{sharePublication}(I_1).\exists \mathsf{sharePublication(B)} \sqcap \exists \mathsf{sharePublication}(I_2).\exists \mathsf{sharePublication(B)}) = 0.65.$$

Because more information was provided the probability inferred was different. The same threshold now would preserve the link.

In order to compare with existing graph-based algorithms, topological features have also been defined. Thus, for every researcher (every node) the number of shared neighbors and possible paths between two nodes have been computed.

4.2. Methodology

In this section we describe our main design choices to run the experiments. According to cross validation principles, our dataset (1100 researchers) has been divided in training and validation sets. To avoid skeweness (due to unbalanced classes,i.e., few links in the Lattes dataset), every fold is comprised by balanced negative and positive instances, where positive instances correspond to a link between two nodes while negative instance means that there is not a link between these two nodes.

In order to classify possible links and therefore to perform comparisons with previous approaches we resort to the Logistic regression classification algorithm.

In a classification approach for link prediction, features are commonly extracted from topological graph properties such as neighbor nodes and paths between nodes. In addition, these numerical features also stem from joint probability distributions and semantics.

To perform comparisons, we resort to two baseline graph-based numerical features. First, the Katz measure [Liben-Nowell and Kleinberg 2003], which is a weighted sum of the number of paths in the graph that connect two nodes, with higher weight for shorter paths. This leads to the following equation:

$$Katz(x, y) = \sum_{i=1}^{\infty} \beta^i p_i$$

where p_i is the number of paths of length i connecting x and y, while β (≤ 1) is a parameter used to regularize this feature. A small value of β considers only shorter paths.

Since computing all paths (∞) is expensive we only consider paths of length at most four ($i \leq 4$).

The second numerical feature is the Adamic-Adar measure [Adamic and Adar 2001] which computes the similarity between two nodes in a graph. Let $\Gamma(x)$ be the set of all neighbors of node x. Then the similarity between two nodes x, y is given by

$$\text{Adamic-Adar}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

The intuition behind the score is that instead of simply counting the number of neighbors shared by two nodes, we should weight the hub nodes less and rarer nodes more. In this way, Adamic-Adar weighs the common neighbors with smaller degree more heavily.

Finally, we also use the probability, $P(r(x, y)|evidence)$, given by our probabilistic description logic model, as a numerical feature in the classification model. We wish to investigate whether this probabilistic logic measure can improve the classification approach for link prediction.

4.3. Results

In order to evaluate suitability of our approach in predicting co-authorships in the Lattes dataset, two experiments were run. In the first experiment two baseline scores, Katz and Adamic-Adar, have been used as features in the logistic regression algorithm. After a ten-fold cross validation process the classification algorithm yielded results on accuracy which are depicted in Table 1.

One can see that on the Lattes dataset, the Katz feature yields the best accuracy (75.495%) when the two baseline features are used in isolation. Katz has been shown to be among the most effective topological measures for the link prediction task [Liben-Nowell and Kleinberg 2003]. Furthermore, when we combine the Katz and the Adamic-Adar features, we improve the accuracy to 75.585%.

Table 1. Classification results on accuracy ($\%$) for baseline features: Adamic-Adar (Adamic), Katz and a combined one (Adamic+Katz)

	Adamic	Katz	Adamic+Katz
Lattes dataset	72.25	75.495	**75.585**

In the second experiment, a probabilistic feature based on our probabilistic description logic approach was introduced into the model. Results on accuracy for this feature are depicted in Table 2. The PDL feature performs better than the other features. This feature alone yields 76.83% on accuracy. When we combine all the three features together, there is an improvement in accuracy to 77.23%.

Table 2. Classification results on accuracy($\%$) for probabilistic description logics and baseline features: CR$\mathcal{ALC}$ based (cralc) and Adamic-Adar, Katz, CR$\mathcal{ALC}$ (Adamic+Katz+cralc).

	cralc	Adamic+Katz+cralc
Lattes dataset	76.83	**77.23**

It is worth noting that the probabilistic logic feature probability outer performs all other features and allow us to improve the classification model for link prediction on accuracy.

Nothing prevent us to define ad-hoc probabilistic networks to estimate link probabilities. However, by doing so we are expected to define a large propositionalized network (a relational Bayesian network) [Revoredo et al. 2011] or estimate local probabilistic networks [Wang et al. 2007]. These approaches do not scale well since computing probabilistic inference for large networks is expensive.

To overcome these performance and scalability issues, we resort to probabilistic inference in CR$\mathcal{ALC}$ which is based on variational methods — tunned by evidence defined

according nodes's neighborhood. Thus, for a ten thousand network, if evidence is given for 5 nodes, then there is only 6 slices which have messages interchanged. In our experiments, the average runtime for inference (1100 nodes network) was 43.401 milliseconds. On the other hand, a propositionalized relational Bayesian network fails to run inference due to out of memory issues.

5. Conclusion

We have presented an approach for predicting links that resorts to both graph-based and ontological information. Given a collaborative network, we encode interests and graph features through a CR$\mathcal{ALC}$ probabilistic ontology. In order to predict links we resort to probabilistic inference, where only information about two nodes being analyzed and the nodes in their path are used as evidence. Thus, making the proposal scalable. Results focused on an academic domain, and we aimed at predicting links among researchers. These preliminary results showed the potential of the idea.

As future work we intend to evaluate other metrics to reduce the number of evidences and consider other datasets.

Previous combined approaches for link prediction [Caragea et al. 2009, Aljandal et al. 2009] have focused on machine learning algorithms [Mitchell 1997]. In such schemes, numerical graph-based features and ontology-based features are computed; then both features are input into a machine learning setting where prediction is performed. Unless from such approaches, in our work we adopt a generic ontology (instead of a hierarchical ontology, expressing only is-a relationships among interests). Therefore, our approach uses more information about the domain to help the prediction.

Acknowledgements

The third author is partially supported by CNPq. The work reported here has received substantial support by FAPESP grant 2008/03995-5 and FAPERJ grant E-26/111484/2010.

References

Adamic, L. A. and Adar, E. (2001). Friends and neighbors on the web. *Social Networks*, 25:211–230.

Aljandal, W., Bahirwani, V., Caragea, D., and Hsu, H. (2009). Ontology-aware classification and association rule mining for interest and link prediction in social networks. In *AAAI 2009 Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0*, Standford,CA.

Baader, F. and Nutt, W. (2002). Basic description logics. In *Description Logic Handbook*, pages 47–100. Cambridge University Press.

Caragea, D., Bahirwani, V., Aljandal, W., and Hsu, W. H. (2009). Ontology-based link prediction in the livejournal social network. In *SARA'09*, pages 1–9.

Cozman, F. G. and Polastro, R. B. (2008). Loopy propagation in a probabilistic description logic. In Greco, S. and Lukasiewicz, T., editors, *Second International Conference on Scalable Uncertainty Management*, Lecture Notes in Artificial Intelligence (LNAI 5291), pages 120–133. Springer.

Cozman, F. G. and Polastro, R. B. (2009). Complexity analysis and variational inference for interpretation-based probabilistic description logics. In *Conference on Uncertainty in Artificial Intelligence*.

Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999). Learning probabilistic relational models. In *Proc. 16th Int. Joint Conference on Artificial Intelligence*, pages 1300–1309.

Heckerman, D., Meek, C., and Koller, D. (2004). Probabilistic entity-relationship models, prms, and plate models. In *In Proceedings of the 21st International Conference on Machine Learning*.

Heinsohn, J. (1994). Probabilistic description logics. In *International Conf. on Uncertainty in Artificial Intelligence*, pages 311–318.

Hsu, W., King, A., Paradesi, M., Pydimarri, T., and Wneinger, T. (2006). Collaborative and structural recommendation of friends using weblog-based social network analysis. In *Proceedings of Computational Approaches to Analysing WebLogs (AAAI)*.

Jaeger, M. (1994). Probabilistic reasoning in terminological logics. In *Principals of Knowledge Representation (KR)*, pages 461–472.

Jaeger, M. (2001). Complex probabilistic modeling with recursive relational bayesian networks. *Annals of Mathematics and Artificial Intelligence*, 32:179–220.

Jaeger, M. (2002). Relational Bayesian networks: a survey. *Linkoping Electronic Articles in Computer and Information Science*, 6.

Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and Knowledge Management*, pages 556–559, New York, NY, USA. ACM.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, New York.

Ochoa-Luna, J., Revoredo, K., and Cozman, F. (2010). Learning sentences and assessments in probabilistic description logics. In *Bobillo, F., et al. (eds.) Proceedings of the 6th International Workshop on Uncertainty Reasoning for the Semantic Web*, volume 654, pages 85–96, Shangai, China. CEUR-WS.org.

Ochoa-Luna, J. E., Revoredo, K., and Cozman, F. G. (2011). Learning probabilistic description logics: A framework and algorithms. In Batyrshin, I. and Sidorov, G., editors, *Advances in Artificial Intelligence*, volume 7094 of *Lecture Notes in Computer Science*, pages 28–39. Springer Berlin Heidelberg.

Polastro, R. B. and Cozman, F. G. (2008). Inference in probabilistic ontologies with attributive concept descriptions and nominals. In *4th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW) at the 7th International Semantic Web Conference (ISWC)*, Karlsruhe, Germany.

Revoredo, K., Ochoa-Luna, J., and Cozman, F. (2010). Learning terminologies in probabilistic description logics. In da Rocha Costa, A., Vicari, R., and Tonidandel, F., editors, *Advances in Artificial Intelligence SBIA 2010*, volume 6404 of *Lecture Notes in Computer Science*, pages 41–50. Springer / Heidelberg, Berlin.

Revoredo, K., Ochoa-Luna, J., and Cozman, F. (2011). Semantic link prediction through probabilistic description logics. In *Bobillo, F., et al. (eds.) Proceedings of the 7th*

International Workshop on Uncertainty Reasoning for the Semantic Web, volume 778, pages 87–97, Bonn, Germany. CEUR-WS.org.

Sachan, M. and Ichise, R. (2011). Using semantic information to improve link prediction results in network datasets. *International Journal of Computer Theory and Engeneering*, 3:71–76.

Sebastiani, F. (1994). A probabilistic terminological logic for modelling information retrieval. In *ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 122–130.

Taskar, B., faiWong, M., Abbeel, P., and Koller, D. (2003). Link prediction in relational data. In *Proceedings of the 17th Neural Information Processing Systems (NIPS)*.

Wang, C., Satuluri, V., and Parthasarathy, S. (2007). Local probabilistic models for link prediction. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 322–331, Washington, DC, USA. IEEE Computer Society.

Wohlfarth, T. and Ichise, R. (2008). Semantic and event-based approach for link prediction. In *Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management*.

Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z. (2006). Stochastic relational models for discriminative link prediction. In *Proceedings of the Neural Information Processing Systems (NIPS)*.