# Assertion Role in a Hybrid Link Prediction Approach through Probabilistic Ontology

**Marcius Armada**[1] **, Kate Revoredo**[1] **, José Eduardo Ochoa Luna**[2] **, Fabio Gagliardi Cozman** [3]

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

[2]Universidad Católica San Pablo
Quinta Vivanco s/n , Urb. Campiña Paisajista, Arequipa, Perú

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

{marcius.oliveira, katerevoredo}@uniriotec.br, eduardo.ol@gmail.com, fgcozman@usp.br

***Abstract.*** *Link prediction in a network is mostly based on information about the neighborhood topology of the nodes. Recently, the interest for hybrid link prediction approaches that combine topology information with information about the network individuals, has grown. However, considering the whole set of individuals may not be necessary and sometimes not even suitable. Therefore, mechanisms to automatically discover the relevant set of individuals are demanding. In this paper, we encompass this problem by proposing an algorithm that combines structure and semantic metrics to find the set of relevant individuals. We empirically evaluate this proposal analyzing the assertion role of these individuals when predicting a link through a probabilistic ontology.*

## 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. These networks have a dynamic behavior, thus nodes and links can appear and disappear rapidly. In this scenario, predicting a possible link in a network, this is predicting a future occurrence of a not yet existing relationship, is an interesting issue that has received significant attention. For instance, one may be interested on finding potential friendship between two persons in a social network, or a potential collaboration between two researchers. In short, *link prediction* aims at predicting whether two nodes should be connected given previous information about their relationships or interests.

Hasan and Zaki [Al Hasan and Zaki 2011] survey representative link prediction methods, classifying them in three groups. In the first group, feature-based methods construct pairwise features to use in classification. The majority of the features are extracted from the graph topology by computing similarity based on the neighborhood of the pair of nodes, or based on ensembles of paths between the pair of nodes [Liben-Nowell and Kleinberg 2007]. Semantic information has also been used as features [Sachan and Ichise 2011, Wohlfarth and Ichise 2008]. The second group includes probabilistic approaches that model the joint probability for entities in a network by Bayesian graphical models [Wang et al. 2007]. The third group employs linear algebraic

approaches that compute the similarity between nodes in a network by rank-reduced similarity matrices [Kunegis and Lommatzsch 2009].

In [Ochoa-Luna et al. 2013], an approach for link prediction that combines Bayesian graphical models and semantic-based features was proposed. To represent semantic-based features, a probabilistic ontology represented with the probabilistic description logic called Credal $\mathcal{ALC}$ (CR$\mathcal{ALC}$) [Cozman and Polastro 2009] was used. This probabilistic description logic extends the popular logic $\mathcal{ALC}$ [Schmidt-Schauß and Smolka 1991] with *probabilistic inclusions*. These are sentences, such as $P(\mathsf{Professor}|\mathsf{Researcher}) = 0.4$, specifying the probability that an element of the domain is a Professor given that it is a Researcher. Exact and approximate inference algorithms for CR$\mathcal{ALC}$ have been proposed [Cozman and Polastro 2009], using ideas inherited from the theory of Relational Bayesian Networks [Jaeger 2002].

When using semantic features, information about the individuals of the domain are considered. However, information about all individuals may not be necessary and sometimes not even suitable. Therefore, mechanisms that automatically select the relevant individuals are important. In [Ochoa-Luna et al. 2013], a first discussion about this matter was done, where structure features were considered to select the most relevant individuals. In this paper, we extend this idea and evaluate alternative methods for selecting the set of relevant individuals. We empirically evaluate our proposal using a probabilistic ontology, represented in CR$\mathcal{ALC}$, for modeling the domain.

The paper is organized as follows. Section 2 reviews basic concepts of probabilistic description logics and link prediction. Our proposal for selecting the most relevant individuals related to the two being analyzed for link prediction is presented in Section 3. Section 4 describes experiments, and Section 5 concludes the paper and discusses some future work.

## 2. Background

This section briefly review probabilistic description logics and link prediction methods, with a focus on concepts and techniques that are later used.

### 2.1. 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. An assertion states that an individual belongs to a concept of that a pair of individuals satisfies a role. An *ABox* is a set of assertions.

A popular description logic is $\mathcal{ALC}$ [Schmidt-Schauß and Smolka 1991]; given its importance to our proposal, we briefly review it here. Constructors in $\mathcal{ALC}$ are *conjunction* ($C \sqcap D$), *disjunction* ($C \sqcup D$), *negation* ($\neg C$), *existential* restriction ($\exists r.C$), and *value* restriction ($\forall r.C$). Concept *inclusions* and *definitions* are denoted respectively by $C \sqsubseteq D$ and $C \equiv D$, where $C$ and $D$ are concepts. Concept $C \sqcup \neg C$ is denoted by $\top$, and concept $C \sqcap \neg C$ is denoted by $\bot$. The semantics of these constructs is given by a domain

$\mathcal{D}$ and an *interpretation* $\mathcal{I}$ as follows: each individual $a$ is mapped into an element $a^{\mathcal{I}}$; each concept $C$ is mapped into a subset $C^{\mathcal{I}}$ of the domain; each role $r$ is mapped into a binary relation $r^{\mathcal{I}}$ in the domain; moreover,

- $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$;
- $(C \sqcup D)^{\mathcal{I}} = C^{\mathcal{I}} \cup D^{\mathcal{I}}$;
- $(\neg C)^{\mathcal{I}} = \mathcal{D} \backslash C^{\mathcal{I}}$;
- $(\exists r.C)^{\mathcal{I}} = \{x \in \mathcal{D} | \exists y : (x, y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$;
- $(\forall r.C)^{\mathcal{I}} = \{x \in \mathcal{D} | \forall y : (x, y) \in r^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}$.

Finally, $C \sqsubseteq D$ is interpreted as $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ and $C \equiv D$ is interpreted as $C^{\mathcal{I}} = D^{\mathcal{I}}$.

An example may be useful. Consider the following concept definition:

$$\mathsf{Researcher} \equiv \mathsf{Person} \ \sqcap \ \exists \mathsf{hasPublication.BibItem} \tag{1}$$

specifying that researchers are individuals who are persons and who have published a bibliographic item.

Several *probabilistic* description logics have appeared in the literature [Lukasiewicz and Straccia 2008, Klinov 2008]. An example is the probabilistic description logic CR$\mathcal{ALC}$ , which is a probabilistic extension of the description logic $\mathcal{ALC}$. 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 $r$ (in this paper we only consider equality in probabilistic inclusions/definitions). If the interpretation of $D$ is the whole domain, then we simply write $P(C) = \alpha$. The semantics of these inclusions is roughly (a formal definition can be found in Ref. [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 (where "use" is the transitive closure of "directly use"; we say that $C$ directly uses $D$ if $D$ appears in the right hand side of an inclusion/definition, or in the conditioning side of a probabilistic inclusion). 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 each value restriction $\forall r.C$ is added to the graph $\mathcal{G}(\mathcal{T})$ as a node, 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.

Consider, as an example, a terminology $\mathcal{T}_R$ containing the sentence in Expression (1), plus $P(\mathsf{Person}) = 0.2$, $P(\mathsf{BibItem}) = 0.6$, $P(\mathsf{hasPublication}) = 0.1$; its graph is depicted in the left of Figure 1.

The semantics of CR$\mathcal{ALC}$ is based on probability measures over the space of interpretations, for a fixed domain. To make sure a terminology specifies a single probability measure, a number of additional assumptions are adopted: the domain is assumed

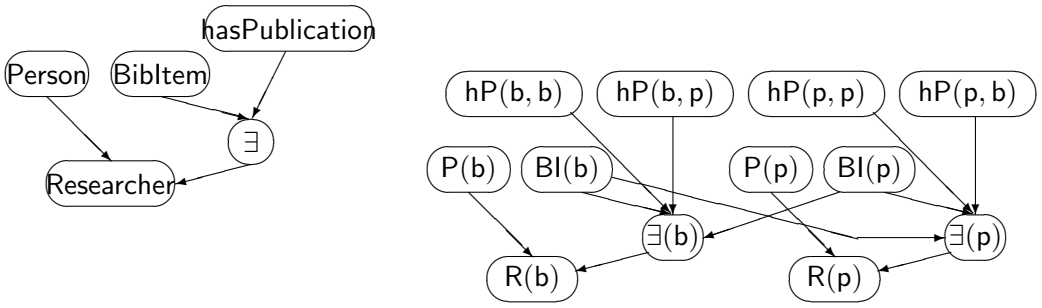


**Figure 1. Graph $\mathcal{G}(\mathcal{T}_R)$ (left Figure) and Bayesian network over indicator functions of assertions, produced by grounding the terminology $\mathcal{T}_R$ (right figure)**

finite, fixed, and known; the unique-name assumption and the rigidity assumption for individuals (as usual in first-order probabilistic logic [Fagin et al. 1990]) are assumed; a single concept name appears in the left hand side of any inclusion or definition and in the conditioned side of any probabilistic inclusion; and finally a Markov condition imposes independence of any grounding of concept/role conditional on the groundings of its corresponding parents in the graph $\mathcal{G}(\mathcal{T})$ [Cozman and Polastro 2009]. Given these assumptions, a set of sentences $\mathcal{T}$ in CR$\mathcal{ALC}$ defines a *relational Bayesian network* [Jaeger 2002] whose underlying graph is exactly $\mathcal{G}(\mathcal{T})$.

Considering the domain $\mathcal{D} = \{\mathsf{bob}, \mathsf{paper}\}$ and the set of assertions $\mathcal{A} = \{$ $\mathsf{Person}(\mathsf{bob}), \mathsf{Researcher}(\mathsf{bob}), \mathsf{BibItem}(\mathsf{paper}), \mathsf{hasPublication}(\mathsf{bob}, \mathsf{paper})$ $\}$, inferences such as $P(\mathsf{A}_o(\mathsf{a}_0)|\mathcal{A})$ can be computed by grounding the terminology, where grounding means that all existing variables must be replaced by constants. In our case they are replaced by the individuals in the domain and the grounding process generates a "slice" for each individual. The right Bayesian network in Figure 1 shows a grounding for terminology $\mathcal{T}$ where two slices, one for individual bob and another for individual paper, are built (for the sake of space, names are abbreviated). At first sight the resulting Bayesian network may seem odd, with nodes like $\mathsf{Bibitem}(\mathsf{bob})$ or $\mathsf{Person}(\mathsf{paper})$, but since we are not based on the "closed world" assumption then anything we not currently known can be either true or false. For large domains, exact probabilistic inference is in general quite hard due to the complexity of the resulting grounded Bayesian network but variational algorithms that approximate such probabilities are available in the literature [Cozman and Polastro 2009] in an attempt to deal whith this problem.

## 2.2. Link Prediction

The task we are interested in can be defined as follows [Liben-Nowell and Kleinberg 2007]. One is given a network (a graph) $G$ consisting of a set of nodes $V$ (represented by letters $a$, $b$, etc) and a set of edges $E$, where an edge represents an interaction between nodes. Interactions may be tagged with times, and the link prediction problem may be one of predicting the existence of edges in a time interval, given the edges observed in another time interval. Here we are interested in a static problem where we are given nodes and edges, except for the edge between two nodes $a$ and $b$, and we must then predict whether there is an edge between $a$ and $b$.

Many different tools are used for link prediction, some of which, like matrix factorization, are related to the massive size of datasets; other tools are directly related to the

existence of links between nodes. One can use classifiers that, based on network features and measures, classify each tentative link as existing or not [Al Hasan and Zaki 2011]; one may also resort to collective classification over the whole set of possible links [Getoor and Diehl 2005]. Several such techniques are based on computing measures of proximity/similarity between nodes in a network [Liben-Nowell and Kleinberg 2007, Lü and Zhou 2011].

Other approaches consider semantic features. The degree of semantic similarity among entities can be useful to predict links that might be missed by simple topological or frequency-based features [Wang et al. 2007]. One way of capturing semantic similarity is by considering documents related to nodes in the network. A simple example of semantic similarity is the keyword match count between two authors [Hasan et al. 2006]. A more sophisticated method makes use of the well-known techniques such as TFIDF feature vector representation and the cosine measure to compute similarity [Wang et al. 2007]. The latter measure, for documents $d_1$ and $d_2$, is obtained by creating vector representations $\overrightarrow{V}(d_1)$ and $\overrightarrow{V}(d_2)$ that countain word counts weighted by their TFIDF (Term Frequency - Inverse Document Frequency) measures. The similarity measure is then

$$\text{cosine}(d_1, d_2) = \frac{\overrightarrow{V}(d_1) \cdot \overrightarrow{V}(d_2)}{|\overrightarrow{V}(d_1)||\overrightarrow{V}(d_2)|},$$

where the dot product is used in the numerator and the Euclidean distance is used in the denominator. To recall, the TFIDF weighting scheme assigns to term $t$ a weight in document $d$ given by $\text{TFIDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t$, where $\text{TF}_{t,d}$ is the term frequency in $d$, and $\text{IDF}_t$ is the inverse document frequency of $t$, given by $\text{IDF}_t = \log \frac{\text{N}}{\text{DF}_t}$, for N the total number of documents and $\text{DF}_t$ the number of documents containing the term.

Approaches to link prediction can be understood not only by considering the kinds of tools employed, but also by examining the model that is used to represent the network as a whole. Typically one assumes some sort of probabilistic mechanism that at least partially explains the existence of edges, perhaps together with domain-specific knowledge (for instance, domain theories about human relationships) [Goldenberg et al. 2010, Newman 2003]. Thus the simplest network model is the Erdös-Rènyi random graph: each pair of nodes can be connected with identical probability. More sophisticated models resort to hierarchical specification of link probabilities, or to grouping of nodes within blocks of varying probability.

One way to capture the probabilistic structure of a network is through graph-based models such as Markov random fields or Bayesian networks [Pearl 1988]. However, these languages are well suited to express independence relations between a fixed set of random variables; when nodes and links are to be dealt within graphs, it is best to consider modeling languages that can specify Markov random fields and Bayesian networks over relational structures. Indeed many proposals for link prediction resort to such languages, from seminal work by Getoor et al [Getoor et al. 2002] and Taskar et al [Taskar et al. 2003]. The presence of relational structure lets one to represent properties of individuals nodes, of links, of communities; one can then compute the probability of specific links, and estimate such probabilities from data.

In [Ochoa-Luna et al. 2013], this modeling strategy was followed using the probabilistic description logics CR$\mathcal{ALC}$. The interest in models based on description log-

ics is justified given recent results on the importance of ontologies in organizing information that can be used in link prediction [Aljandal et al. 2009, Caragea et al. 2009, Thor et al. 2011]. While other link prediction implementation usually focus in one kind of feature, the one using CR$\mathcal{ALC}$ showed to be able to mix different features such as semantic, numeric and topological. Being a versatile solution doesn't make it easier to be modeled than other solutions, but as a novel approach there is still room for evolution and further experimentation.

## 3. Assertion Role in Link Prediction through a Probabilistic Ontology

Given a network (a graph) $G$ consisting of a set of nodes $V$ and a set of edges $E$, where an edge represents an interaction between nodes. For a link prediction task considering semantic features, we follow the approach proposed in [Ochoa-Luna et al. 2013] and model the domain using a probabilistic ontology ($O$) represented in CR$\mathcal{ALC}$. Nodes in $G$ are individuals of a concept $C$ in $O$ and edges are instances of a role $R$ in $O$. Thus, the network G is built encompassing assertions about concept $C$ and role $R$. For instance, in a co-authorship network, assertions for concept Researcher are represented by nodes and assertions for role sharePublication are represented by relationships between two nodes. Figure 2 depicts a network for the assertions shown in Figure 3.

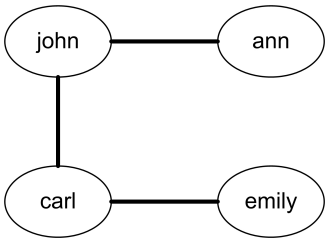


**Figure 2. Network encompassing assertions of the ABox in Figure 3.**

Researcher(john). Researcher(ann). Researcher(carl).
Researcher(emily). sharePublication(john, ann).
sharePublication(john, carl). sharePublication(carl, emily).

**Figure 3. Example of an ontology ABox.**

The probabilistic ontology $O$ can model the domain widely, thus having other concepts and roles beyond the ones encompassing the network. For instance, an ontology describing the co-authorship domain is shown in Figure 4.

TBox:

| | | |
|---|---|---|
| $P$(Publication) | = | 0.3 |
| $P$(sharePublication) | = | 0.22 |
| $P$(hasSameInstitution) | = | 0.14 |
| Researcher | ≡ | Person ⊓ ∃hasPublication.BibItem |
| $P$(PublicationCollaborator | \| | Researcher ⊓ ∃sharePublication.Researcher) = 0.91 |

ABox:

Researcher(john). Researcher(ann). Researcher(carl).
Researcher(emily). sharePublication(john, ann).
sharePublication(john, carl). sharePublication(carl, emily).
Publication(p1). Publication(p2)

**Figure 4. A probabilistic ontology for the co-authorship domain.**

Predicting a link between two nodes $a$ and $b$ in a network G concerns evaluating whether an edge between $a$ and $b$ should be included. In the semantic link prediction task,

where the domain is modeled through CR$\mathcal{ALC}$, the problem can be rewritten as evaluating if the considered role between individuals $a$ and $b$ may exist in a given ontology. Thus, the semantic link prediction task considered in this paper can be described as: compute the probability of an assertion concerning the role that provides the semantic of relationships in the network G given an ABox of asserted concepts and roles of the domain.

Because domain knowledge is expressed with CR$\mathcal{ALC}$, questions about probability of assertions can be answered by inference in CR$\mathcal{ALC}$. For instance, the question "what is the probability of Emily and Ann share a publication given some information about the domain?" can be translated into $P(\mathsf{sharePublication}(\mathsf{emily}, \mathsf{ann})|\mathcal{A})$, where $\mathcal{A}$ represents the ABox with assertions about the domain. If this probability is higher than a suitable threshold then the assertion may be considered true and a link introduced in G.

Intuitively, the inference quality of any assertion's probability rests in the used assertions contained in $\mathcal{A}$. While one can suppose that more assertions leads to more accurate calculated probabilities, this is not always true. Some individuals may not be related to the ones being analyzed and therefore their assertions may not impact the evaluation. Thus it is unnecessary to consider evidence (assertion) about them. Moreover, in some case may even be impractical to reason about all individuals of the domain due to limits in computational resources or long response times. Hence it is important to filter out assertions and to focus on the most relevant ones.

We are interested in predicting a relationship between two specific nodes, $a$ and $b$. Therefore, we argue that assertions directly related to these two individuals, and to other individuals strongly related to them in the network, are more relevant for link prediction than assertions on other individuals in the network. The link prediction algorithm (see Algorithm 1) will not only be scalable but will be more accurate if we only consider assertions about $a$, $b$ and the individuals strongly related to them in our inferences. To do so, we must specify the set $\mathcal{A}(a, b)$ of elements of the domain that are deemed strongly related to $a$ and $b$.

**Algorithm 1:** Algorithm for link prediction (adapted from [Ochoa-Luna et al. 2013]).

**Require:** a network $G$, an ontology $\mathcal{O}$, a role $\hat{r}$ representing links in the network, a concept $\hat{C}$ specifying the nodes in the network and a threshold $\gamma$.
**Ensure:** a set of predicted links $L$

```
initialize L = ∅;
for all pair of instances (a, b) of nodes in G do
    if there is no link between nodes a and b in G then
        find A(a, b);
        E=assertions about A(a, b);
        infer probability P(r(a, b)|E) using the relational Bayesian network created
        from the ontology O;
        if P(r(a, b)|E) > γ then
            add link between a and b to L.
        end if
    end if
end for
```

In [Ochoa-Luna et al. 2013] the strategy adopted to define $\mathcal{A}(a, b)$ was to consider nodes along paths between $a$ and $b$. In this paper, we argue that not only structural metrics can define the best set $\mathcal{A}(a, b)$ and we evaluate the performance of structural and semantic approaches for selecting the most relevant individuals for a link prediction task. The following approaches were considered:

i) $\mathcal{A}(a, b) = \mathcal{A}_{adj}(a, b)$, where $\mathcal{A}_{adj}(a, b) = adjacent(a) \cup adjacent(b)$. Defines $\mathcal{A}(a, b)$ as the set of nodes adjacent to $a$ union the set of nodes adjacent to $b$.
ii) $\mathcal{A}(a, b) = \mathcal{A}_{Padj}(a, b)$, where $\mathcal{A}_{Padj}(a, b) = A_0(a, b) \cup_{i \in \mathcal{A}_0(a,b)} adjacent(i)$ and $\mathcal{A}_0(a, b) = \{a\} \cup \{b\} \cup path(a, b)$. Defines $\mathcal{A}(a, b)$ as the set of all nodes in the path between $a$ and $b$ union their adjacent nodes and the adjacents of $a$ and $b$.
iii) $\mathcal{A}(a, b) = f_{semantic}(\mathcal{A}_{adj}(a, b))$. Defines $\mathcal{A}(a, b)$ as the set of nodes contained in $\mathcal{A}_{adj}(a, b)$ that are most semantically related to $a$ and $b$ considering a semantic function $f_{semantic}$.
iv) $\mathcal{A}(a, b) = f_{semantic}(\mathcal{A}_{Padj}(a, b))$. Defines $\mathcal{A}(a, b)$ as the set of nodes contained in $\mathcal{A}_{Padj}(a, b)$ that are most semantically related to $a$ and $b$ considering a semantic function $f_{semantic}$.

An experimental evaluation was conducted and will be described in the next section to evaluate the benefits of these metrics. Moreover, a discussion around the role of the assertions about individuals for the semantic link prediction task is also presented.

## 4. Experiments

Experiments have been conducted to evaluate the benefits of considering structural and semantic metrics for selecting the most relevant individuals for the semantic link prediction task. A real world data repository, the Lattes curriculum platform, was used. This section reports the steps involved in this process and the results found.

### 4.1. Scenario Description

The Lattes platform is the public repository of brazilian scientific curricula that consists of approximately a million registered documents. Information is encoded in HTML format, ranging from personal information such as name and professional address to publication lists, administrative tasks, research areas, research projects and advising/advisor information. There is implicit relational information in these web pages, for instance collaboration networks are built by advising/adviser links, shared publications, and so on.

To perform experiments we have randomly selected eight thousand researchers and their relationships from the Lattes platform. Assertions were extracted concerning these researchers. For instance, if a parser finds that a researcher John has two publications ($p_1$, $p_2$) and a researcher Ann has two ($p_2$, $p_3$), where $p_2$ was done in collaboration with John, then assertions, as the following, are extracted:

$\mathsf{Researcher(john)}$, $\mathsf{Researcher(ann)}$,
$\mathsf{Publication(p_1)}$, $\mathsf{Publication(p_2)}$, $\mathsf{Publication(p_3)}$,
$\mathsf{hasPublication(john, p_1)}$, $\mathsf{hasPublication(john, p_2)}$,
$\mathsf{hasPublication(ann, p_2)}$, $\mathsf{hasPublication(ann, p_3)}$
$\mathsf{sharePublication(john, ann)}$.

A probabilistic ontology was then learned using algorithms in the literature [Ochoa-Luna et al. 2011, Revoredo et al. 2010]. This ontology is comprised by 24 probabilistic inclusions and 17 concept definitions.

The concept Researcher indicates whether an element of the domain is a node in the network (hence for each assertion of concept Researcher a node exists in the network) and the role sharePublication indicates whether a pair of elements of the domain are linked in the network (hence for each assertion of role sharePublication a link exists in the network). Using this data, link probabilities were computed through inference in the CR$\mathcal{ALC}$ ontology.

### 4.2. Methodology

In this section, we describe our main design choices to run experiments. Given the 8000 selected researchers, there exist $31996000$ possible link relationships. To perform link prediction we have considered collaborations based on co-authorship on publications (there are $2837206$ publications). After analysing these publications we identified $95011$ true positive links among researchers based on co-authorship. From the available data we randomly selected links so that the used dataset in the experiments was comprised by 1000 positive links and 1000 negative links (balanced datasets).

Although we can use probabilistic inference to decide whether there is a link between two nodes, to perform comparisons among the structural and semantic metrics described in Section 3 we resort to a classification algorithm approach through the Logistic regression algorithm.

Beyond the 4 metrics described in Section 3 we also considered:

v) the metric proposed in [Ochoa-Luna et al. 2013]: $\mathcal{A}(a,b) = \mathcal{A}_{path}(a,b)$, where $\mathcal{A}_{path}(a,b)$ defines the set of nodes in the paths between $a$ and $b$ .
vi) $\mathcal{A}(a,b) =$ random selection of $10$ nodes in the network.

The metric v will permit us compare our proposal with the previous one presented in [Ochoa-Luna et al. 2013]. For this metric, since computing all paths ($\infty$) is expensive, we follow Ochoa et al. and only considered paths of length at most four ($i \leq 4$).

The semantic feature we considered was keyword match. For each researcher a document with the words appearing in the title of his publications (removing stop words) is considered. Thus, a researcher is represented as a set of words, which allows us to compute a semantic feature: the keyword *match* count between two researchers [Hasan et al. 2006]. Using this feature we were able to select the top 10 researchers with the most words in common with $a$ an $b$.

Finally, the probability $P(r(x,y)|E)$, given by our probabilistic description logic model, is used as a numerical feature in the classification model, in order to investigate whether it 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, several experiments were conducted. Each metric, through the probabilistic logic scores found, has been considered as isolated features in our clasification algorithm. After

**Table 1. Classification results for dataset Lattes on accuracy (%) for baseline features used for selecting individuals used for generating assertions for inference in CR$\mathcal{ALC}$ : metric i, metric ii, metric iii, metric iv, metric v, metric vi.**

| Feature | Lattes (acc.) | Avg(#) of selected individuals |
|---|---|---|
| CR$\mathcal{ALC}$ + metric i | 99.93% | 501 |
| CR$\mathcal{ALC}$ + metric ii | 99.86% | 545 |
| CR$\mathcal{ALC}$ + metric iii | 99.88% | 10 |
| CR$\mathcal{ALC}$ + metric iv | 99.65% | 10 |
| CR$\mathcal{ALC}$ + metric v | 92.41% | 26 |
| CR$\mathcal{ALC}$ + metric vi | 71.14% | 10 |

a ten-fold cross validation process, the classification algorithm yielded results on accuracy for the dataset which are depicted in Table 1.

The results shows us that randomly selecting individuals for assertion generation (metric vi) obtained the worse accuracy in comparison to the other metrics with only 71% while all the other obtained accuracies greater than 90%. Thus, it is important to use the best possible assertions in the inference.

All other results show little differences in accuracy between each other but those metrics which don't use the semantic feature (metric i and ii) needed about 50 times more individuals to obtain near the same results. This demonstrates that the quality of the selected individuals, using the semantic feature, and the assertions generated from them were able to keep the CR$\mathcal{ALC}$ link prediction algorithm scalable and the quality of the predictions high.

## 5. Conclusion

In this paper, we have evaluated the role of assertions about individuals for the semantic link prediction task. We follow the approach introduced in [Ochoa-Luna et al. 2013] and considered a probabilistic ontology, represented with the probabilistic description logic CR$\mathcal{ALC}$, for modeling the domain. Thus, given a collaborative network, interests and graph features are encoded through the probabilistic ontology.

To predict links, probabilistic inference is used. Structural and semantic metrics are combined in order to select the most relevant individuals for the prediction link task. Therefore, only the necessary individuals are used and results have shown the importance of selecting the best individuals from the avaiable ones. Moreover, this approach makes the proposal scalable. Our proposal was evaluated on an academic domain, where links among researchers were predicted and was able to attain accuracies greater than 90% as shown in Table 1.

Compared to previous work, our approach employs a rich ontology (as opposed to simple is-a terminologies) that can encode substantial information about the domain. Hierarchical structure can be encoded together with knowledge about specific nodes in a network — we plan to explore richer ontologies in the future. Our proposal attains better scalability than previous proposals that have tried to explore probabilistic relational models for similar purposes but we plan to experiment with other new and state-of-the-

art selection algorithms in the search for the best set of assertions to be used in the link prediction task.

## 6. Acknowledgment

This work is being accomplished in the context of the "Infrastructure for the Management of Scientific Experiments in Computational Modeling" project, granted by CNPq, No. 559998/2010-4.

## References

Al Hasan, M. and Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics*, pages 243–275. Springer.

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 1–1.

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*.

Fagin, R., Halpern, J. Y., and Megiddo, N. (1990). A logic for reasoning about probabilities. *Information and Computation*, 87:78–128.

Getoor, L. and Diehl, C. P. (2005). Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12.

Getoor, L., Friedman, N., Koller, D., and Taskar, B. (2002). Learning probabilistic models of link structure. *Journal of Machine Learning Research*, 3:679–707.

Goldenberg, A., Fienberg, S. E., Zheng, A. X., and Airoldi, E. M. (2010). A survey of statistical network models.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

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

Klinov, P. (2008). Pronto: A non-monotonic probabilistic description logic reasoner. In *The Semantic Web: Research and Applications*, pages 822–826. Springer.

Kunegis, J. and Lommatzsch, A. (2009). Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 561–568. ACM.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170.

Lukasiewicz, T. and Straccia, U. (2008). Managing uncertainty and vagueness in description logics for the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):291–308.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

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.

Ochoa-Luna, J. E., Revoredo, K., and Cozman, F. G. (2013). Link prediction using a probabilistic description logic. *Journal of the Brazilian Computer Society*, pages 1–13.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference*. Morgan Kaufman.

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.

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.

Schmidt-Schauß, M. and Smolka, G. (1991). Attributive concept descriptions with complements. *Artificial intelligence*, 48(1):1–26.

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

Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., and Zhang, X.-N. (2011). Link prediction for annotation graphs using graph summarization. In *The Semantic Web–ISWC 2011*, pages 714–729. Springer.

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*.