J Braz Comput Soc (2013) 19:397–409
DOI 10.1007/s13173-013-0108-8

ORIGINAL PAPER

Link prediction using a probabilistic description logic

**José Eduardo Ochoa Luna · Kate Revoredo ·
Fabio Gagliardi Cozman**

Received: 23 November 2012 / Accepted: 7 March 2013 / Published online: 25 April 2013

Abstract Due to the growing interest in social networks, link prediction has received significant attention. Link prediction is mostly based on graph-based features, with some recent approaches focusing on domain semantics. We propose algorithms for link prediction that use a probabilistic ontology to enhance the analysis of the domain and the unavoidable uncertainty in the task (the ontology is specified in the probabilistic description logic $\textsc{cr}\mathcal{ALC}$). The scalability of the approach is investigated, through a combination of semantic assumptions and graph-based features. We evaluate empirically our proposal, and compare it with standard solutions in the literature.

Keywords Link prediction · Probabilistic logic · Description logics

1 Introduction

Many social, biological, and information systems can be well described as 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 received significant attention. For instance, one may be interested in finding potential friendships 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.

Mohammad and Mohammed [18] survey representative link prediction methods, classifying them into 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 [15]. Semantic information has also been used as features [26,32]. The second group includes probabilistic approaches that model the joint probability for entities in a network by Bayesian graphical models [31]. The third group employs linear algebraic approaches that compute the similarity between nodes in a network by rank-reduced similarity matrices [14].

We present an approach for link prediction that combines Bayesian graphical models and semantic-based features. Hence, our proposal belongs to the first two categories mentioned in the previous paragraph. To represent semantic-based features, we employ a probabilistic description logic called Credal $\mathcal{ALC}$ ($\textsc{cr}\mathcal{ALC}$) [5]. This probabilistic description logic extends the popular logic $\mathcal{ALC}$ [27] 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 $\mathsf{Professor}$ given that it is a $\mathsf{Researcher}$. Exact and approximate inference algorithms for $\textsc{cr}\mathcal{ALC}$ have been proposed [5], using ideas inherited from the theory of Relational Bayesian Networks [12]. We benefit from such algorithms, and add some techniques to make our approach scalable to real domains. We also present experimental validation of our proposal.

J. E. Ochoa Luna · F. G. Cozman
Escola Politécnica, Universidade de São Paulo,
Av. Prof. Mello Morais 2231, São Paulo, SP, Brazil
e-mail: eduardo.ol@gmail.com

F. G. Cozman
e-mail: fgcozman@usp.br

K. Revoredo (✉)
Departamento de Informática Aplicada, Unirio,
Av. Pasteur, 458, Rio de Janeiro, RJ, Brazil
e-mail: katerevoredo@uniriotec.br

The paper is organized as follows. Section 2 reviews basic concepts of probabilistic description logics and of link prediction. Our proposals for a scalable semantic link prediction approach appear in Sect. 3. Section 4 describes experiments, and Sect. 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) [3]. 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 $\mathcal{ALC}$ [27]; 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:

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

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

Several *probabilistic* description logics have appeared in the literature [13,17]; here we just indicate a few representative proposals.

Heinsohn [11] and Sebastiani [28] consider 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* semantics: 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. The latter semantics characterizes an *interpretation-based* semantics.

The probabilistic description logic CR$\mathcal{ALC}$ is a probabilistic extension of the description logic $\mathcal{ALC}$ that adopts an interpretation-based semantics. 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. [5]) 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 $P(\mathsf{Person}) =$

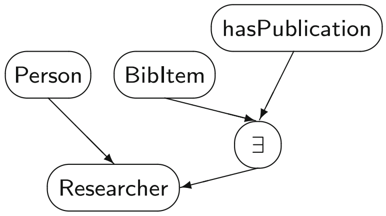

Fig. 1 Graph $\mathcal{G}(\mathcal{T}_R)$

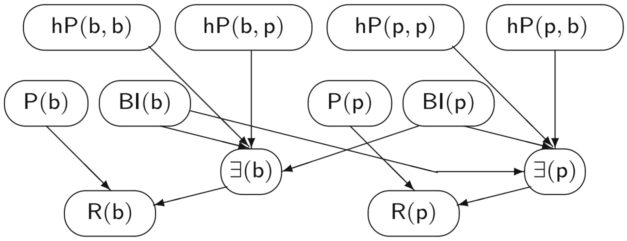

Fig. 2 Bayesian network over indicator functions of assertions, produced by grounding the terminology $\mathcal{T}_R$

0.2, $P(\mathsf{BibItem}) = 0.6$, $P(\mathsf{hasPublication}) = 0.1$; its graph is depicted in Fig. 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 finite, fixed, and known; the unique-name assumption and the rigidity assumption for individuals (as usual in first-order probabilistic logic [6]) 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 $\mathcal{G}(\mathcal{T})$ [5]. Given these assumptions, a set of sentences $\mathcal{T}$ in CR$\mathcal{ALC}$ defines a *relational Bayesian network* [12] whose underlying graph is exactly $\mathcal{G}(\mathcal{T})$.

Consider the following example. Suppose we have terminology $\mathcal{T}_R$ and domain $\mathcal{D} = \{\mathsf{bob}, \mathsf{paper}\}$, There are several possible sets of assertions that are obtained by grounding. For instance,

$$\{\mathsf{Person}(\mathsf{bob}), \mathsf{Researcher}(\mathsf{bob}),$$

$$\mathsf{BibItem}(\mathsf{paper}), \mathsf{hasPublication}(\mathsf{bob}, \mathsf{paper})\}.$$

The assumptions discussed in the previous paragraph induce a single probability measure over the set of all assertions (groundings), because they induce a Bayesian network over indicator variables of assertions.

For example, for domain $\mathcal{D} = \{\mathsf{bob}, \mathsf{paper}\}$, Fig. 2 depicts the Bayesian network over indicator variables of assertions (for the sake of space, names are abbreviated; for instance, hP denotes $\mathsf{hasPublication}$; b denotes bob, and so on). To simplify notation, the indicator function of assertion $C(a)$ is indicated simply by $C(a)$, instead of the more usual convention $I_{C(a)=\mathsf{true}}$.

Inferences, such as $P(\mathsf{A}_o(\mathsf{a}_0)|\mathcal{A})$ for an ABox $\mathcal{A}$, can be computed by grounding, thus generating a Bayesian network where one "slice" is built for each individual. For instance, in the Bayesian network depicted in Fig. 2 two slices, one for individual bob and another for individual paper, are built. For large domains, exact probabilistic inference is in general quite hard. Variational algorithms that approximate such probabilities are available in the literature [5].

2.2 Link prediction

The task we are interested in can be defined as follows [15]. 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 [18]; one may also resort to collective classification over the whole set of possible links [7]. Several such techniques are based on computing measures of proximity/similarity between nodes in a network [15,16]. One of them is the Katz measure [15], a weighted sum of the number of paths in the graph between two given nodes, with higher weights assigned to shorter paths:

$$\mathrm{Katz}_\beta(A, B) = \sum_{i=1}^{\infty} \beta^i p_i,$$

where p_i is the number of paths of length i connecting A and B, while $\beta \in (0, 1]$ weighs the paths—a small value of β favors shorter paths. Another notable proximity measure is the Adamic–Adar measure [1], given by:

$$\mathrm{Adamic{-}Adar}(A, B) = \sum_{C \in \Gamma(A) \cap \Gamma(B)} \frac{1}{\log |\Gamma(C)|},$$

where $\Gamma(X)$ be the set of all neighbors of node X. The intuition behind the Adamic–Adam measure is that, instead of simply counting the number of neighbors shared by two nodes, we should emphasize common neighbors that have less neighbors themselves.

Other approaches to link prediction 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 [31]. 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 [10]. A more sophisticated method makes use of the well-known techniques such as TFIDF feature vector representation and the cosine measure to compute similarity [31]. 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 contain 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 $\text{TFIDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_\text{t}$, where $\text{TF}_{t,d}$ is the term frequency in d, and IDF_t is the inverse document frequency of t, given by $\text{IDF}_\text{t} = \log \frac{N}{\text{DF}_\text{t}}$, for N the total number of documents and 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) [9,19]. 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 [23]. 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. [8] and Taskar et al. [29]. 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 this paper, we follow this modeling strategy; the difference between our modeling language and previous proposals is that we adopt a language based on description logics, as already indicated in the previous section. Our interest in models based on description logics is justified given recent results on the importance of ontologies in organizing information that can be used in link prediction [2,4,30].

3 Link prediction with $\text{CR}\mathcal{ALC}$

Given a network G where many links are observed, one is interested in predicting whether a link between nodes a and b exists (presumably the linkage between a and b has not been observed). We address this problem by considering, in addition to topological information about the network, knowledge about the domain concerning network entities. To do so, domain knowledge is represented through a probabilistic ontology using $\text{CR}\mathcal{ALC}$. Among the concepts (N_C) and roles (N_R) in the ontology, there is a concept $\hat{C}$ that indicates which elements of the domain are nodes in G, and a role $\hat{r}$ that indicates which pairs of elements are linked—hence $\hat{C}$ and $\hat{r}$ describe the network itself, while other concepts and roles describe the remaining domain knowledge. In our experience, it is important to explicitly indicate which elements of the domain are nodes, to make sure inference runs only with the required elements (in effect this is providing a type that separates network nodes from other elements of the domain).

For example, in a coauthorship network, nodes represent researchers and relationships may be “has a publication with” or “is advised by”. An ontology for such a domain, represented by $\text{CR}\mathcal{ALC}$, is shown in Fig. 3. The ontology describes publications, using concepts such as Researcher and Publication, and using roles such as hasPublication, hasSameInstitution, sharePublication. Nodes in the network instantiate a concept (for instance Researcher), while links in the network instantiate a role (for instance sharePublication).

The semantic link prediction task proposed in this paper can be described as: compute the probability of an assertion concerning a particular role of interest, given an ABox $\mathcal{A}$ of asserted concepts and roles involving nodes in the network. Because domain knowledge is expressed with $\text{CR}\mathcal{ALC}$, questions about probability of assertions can be answered by inference in $\text{CR}\mathcal{ALC}$. 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

$$
\begin{array}{ll}
P(\mathsf{Publication}) & = 0.3 \\
P(\mathsf{sharePublication}) & = 0.22 \\
P(\mathsf{hasSameInstitution}) & = 0.14 \\
\mathsf{Researcher} & \equiv \mathsf{Person} \ \sqcap \ \exists \mathsf{hasPublication.BibItem} \\
P(\mathsf{PublicationCollaborator} \mid & \mathsf{Researcher} \sqcap \exists \mathsf{sharePublication.Researcher}) = 0.91
\end{array}
$$

$$
\begin{array}{ll}
\text{ABox:} & \mathsf{Researcher(john)}. \ \ \mathsf{Researcher(ann)}. \ \ \mathsf{Researcher(carl)}. \ \ \mathsf{Researcher(emily)}. \\
 & \mathsf{sharePublication(john, ann)}. \ \ \mathsf{sharePublication(john, carl)}. \ \ \mathsf{sharePublication(carl, emily)}.
\end{array}
$$

Fig. 3 A probabilistic ontology for the co-authorship domain, and an ABox

```
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 $L = \emptyset$;
initialize $E$ = evidence (set of all assertions);
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
    infer probability $P(r(a, b)|E)$ using the relational Bayesian network created from the ontology $\mathcal{O}$;
    if $P(r(a, b)|E) > \gamma$ then
      add link between $a$ and $b$ to $L$.
    end if
  end if
end for
```

Algorithm 1: Algorithm for link prediction: evidence is the complete set of assertions

the information about the domain. If this probability is higher than a suitable threshold, then a link is included.

Our first link prediction algorithm is described in Algorithm 1.[1]

The algorithm starts by going through all pairs of instances of the concept $\hat{C}$ (that is, all nodes). For each pair, it checks whether a link between the corresponding nodes exist in the network; if not, the probability of the link is computed using the relational Bayesian network extracted from the ontology $\mathcal{O}$. If the probability is greater than a threshold, then the corresponding link is added to the set of suggested links. (Alternatively, when the threshold is not given, a list of links, ranked by their probability, can be produced.)

The evidence is the given set of assertions; the size of this set has great impact in inference effort. When inferences are computed, the ontology is turned into a relational Bayesian network, whose grounding is a Bayesian network—each assertion may generate a new slice of nodes in this grounded Bayesian network. Approximate algorithms are necessary for inference; in this work we employ the variational inference method described in Ref. [5]. While one can suppose that more assertions lead to more accurate predictions, the computational effort involved in inference may be so large as to generate bad approximations. 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, assertions directly related to these two objects and to other objects strongly related to them in the network are more relevant for link prediction than assertions on other objects in the network. We can make our link prediction algorithm scalable if we only consider assertions about a, b and about the objects 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.

Liben-Nowell and Kleinberg [15] compute similarities between two nodes using ensembles of paths between the two nodes (so as to decide whether to include a link between the nodes). It seems reasonable to adopt the same strategy, and define $\mathcal{A}(a, b)$ to contain nodes in paths between a and b (although we could consider all possible paths between two nodes, compute this could be expensive. Hence, we restrict ourselves to a path size of five). Therefore, in Algorithm 1 the evidence must be specialized for each pair of nodes; given a and b, the set $\mathcal{A}(a, b)$ must be constructed and the relevant assertions are then collected into E.

The resulting link prediction algorithm is described in Algorithm 2. Experiments with this algorithm, using real data, are reported in the next section.

[1] This algorithm was first discussed in Ref. [25], and later refined, together with Algorithm 2, in Refs. [22] and [21]; the presentation is here further refined. Some experiments and results reported here appeared in those preliminary publications; in this paper we also describe novel experiments with significantly larger datasets.

```
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 $L = \emptyset$;
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
        initialize $E$ = evidence in $\mathcal{A}(a, b)$;
        infer probability $P(r(a, b)|E)$ using the relational Bayesian network created from the ontology $\mathcal{O}$;
        if $P(r(a, b)|E) > \gamma$ then
            add link between $a$ and $b$ to $L$.
        end if
    end if
end for
```

Algorithm 2: Algorithm for link prediction: evidence on nodes and strongly related elements of the domain

4 Experiments

Experiments have been conducted to evaluate our approach to semantic link prediction. A real world data repository, the Lattes curriculum platform, was used. Our algorithm was combined with state-of-the-art classifiers for link prediction. This section reports the steps involved in this process.

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 that are associated with eight research areas. Table 1 depicts these research areas.

Table 1 Research areas and number of co-authored collaboration

Research area	Code	*Number*
Agricultural Sciences	A1	17,157
Biological Sciences	A2	23,222
Exact and Earth Sciences	A3	18,440
Human Sciences	A4	2,281
Social Sciences	A5	4,462
Health Sciences	A6	17,255
Engineering	A7	10,879
Languages and Arts	A8	1,315

Assertions were extracted from the Lattes platform concerning these researchers. For instance, if a parser finds that a researcher John has four publications (p_1, p_2, p_3, p_4) and a researcher Mary has two (p_2, p_5), where p_2 was done in collaboration with John, then assertions, as the following, are extracted:

$\mathsf{Researcher(john)}$, $\mathsf{Researcher(ann)}$,
$\mathsf{Publication(p_1)}$, $\mathsf{Publication(p_2)}$, $\mathsf{Publication(p_3)}$,
$\mathsf{Publication(p_4)}$, $\mathsf{Publication(p_5)}$
$\mathsf{sharePublication(john, ann)}$.

A probabilistic ontology was then learned using algorithms in the literature [20,24]. This ontology is comprised by 24 probabilistic inclusions and 17 concept definitions. Because learning is mainly concerned with deterministic and probabilistic inclusions, the learned ontology was enlarged with 4 relevant roles. Parts of the final ontology can be seen in Figs. 3 and 4.

In this probabilistic ontology, concepts and probabilistic inclusions typically denote mutual research interests. In short, in this ontology a ResearcherLattes is a person that has publications, advises other people and participates on examination boards. On the other hand, a SupervisionCollaborator is a probabilistic inclusion which denotes a kind of researcher that was advised for another researcher. The SameInstitution concept denotes researchers that work at the same institution. Seemingly, the SameBoard concept denotes researchers that have participated on same examination boards. The NearCollaborator is a probabilistic inclusion that denotes researchers working at the same institution that have shared publications. The FacultyNearCollaborator is a near collaborator that also participates of same examination boards. The NullMobility Researcher concept denotes researchers which have low mobility, i.e., they remain at the same institution where they were advised. The StrongRelatedResearcher denotes strong relationship between two researchers (advisor and advisee) which also share publications.

The concept Researcher indicates whether an element of the domain is a node in the network (hence Researcher is $\hat{C}$) and the role sharePublication indicates whether a pair of elements of the domain is linked in the network (hence sharePublication is $\hat{r}$).

$$
\begin{array}{lll}
P(\mathsf{Board}) & & = 0.33 \\
P(\mathsf{wasAdvised}) & & = 0.05 \\
P(\mathsf{sameExaminationBoard}) & & = 0.31 \\
\mathsf{ResearcherLattes} & \equiv & \mathsf{Person} \sqcap (\exists \mathsf{hasPublication.Publication} \ \sqcap \exists \mathsf{advises.Person} \sqcap \exists \mathsf{participate.Board}) \\
P(\mathsf{SupervisionCollaborator} & | & \mathsf{Researcher} \sqcap \exists \mathsf{wasAdvised.Researcher}) = 0.94 \\
P(\mathsf{SameInstitution} & | & \mathsf{Researcher} \sqcap \exists \mathsf{hasSameInstitution.Researcher}) = 0.92 \\
P(\mathsf{SameBoard} & | & \mathsf{Researcher} \sqcap \exists \mathsf{sameExaminationBoard.Researcher}) = 0.95 \\
P(\mathsf{NearCollaborator} & | & \mathsf{Researcher} \sqcap \exists \mathsf{sharePublication}.\exists \mathsf{hasSameInstitution}.\exists \mathsf{sharePublication.Researcher}) = 0.95 \\
\mathsf{FacultyNearCollaborator} & \equiv & \mathsf{NearCollaborator} \ \sqcap \ \exists \mathsf{sameExaminationBoard.Researcher} \\
P(\mathsf{NullMobilityResearcher} & | & \mathsf{Researcher} \sqcap \exists \mathsf{wasAdvised}.\exists \mathsf{hasSameInstitution.Researcher}) = 0.98 \\
\mathsf{StrongRelatedResearcher} & \equiv & \mathsf{Researcher} \ \sqcap \ (\exists \mathsf{sharePublication.Researcher} \ \sqcap \exists \mathsf{wasAdvised.Researcher}) \\
\mathsf{InheritedResearcher} & \equiv & \mathsf{Researcher} \ \sqcap \ (\exists \mathsf{sameExaminationBoard.Researcher} \ \sqcap \exists \mathsf{wasAdvised.Researcher})
\end{array}
$$

Fig. 4 A probabilistic ontology CR$\mathcal{ALC}$ for the Lattes domain

Topological graph information was computed using the assertions for Researcher and sharePublication. Figures 5, 6 and 7 depict collaboration networks within research areas in our dataset.

Using this data, link probabilities were computed through inference in the CR$\mathcal{ALC}$ ontology. To illustrate inference, consider Fig. 8 which depicts a subset of collaborations among researchers. If we inspect this collaboration graph we could be interested, for instance, in checking links among researchers from different groups. Since filling forms in the Lattes platform is prone to errors, there is uncertainty regarding real collaborations. Thus, in Fig. 8 one could further investigate whether a link between researcher R (rectangle node) and the researcher B (triangle node) is suitable. The probability of a possible link between R and B was computed, $P(\mathsf{sharePublication}(R, B)|E)$, where E contains evidence about researchers such as publications, institution, examination board participations and so on. Using evidence

$$\mathsf{Researcher}(R) \sqcap \exists \mathsf{hasSameInstitution.Researcher}(B),$$

one obtains

$$P(\mathsf{sharePublication(R, B)|Researcher(R)} \\ \sqcap \exists \mathsf{hasSameInstitution.Researcher}(B)) = 0.57.$$

One could obtain more evidence, such as information about nodes that indirectly connect these two groups (Fig. 8), denoted by I_1, I_2. Consider:

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

In order to compare our approach with existing algorithms, topological and semantic features have also been defined, as discussed in the following sections.

4.2 Methodology

In this section, we describe our main design choices to run experiments.

Given the 8,000 selected researchers, there exist 31,996,000 possible link relationships. To perform link prediction we have considered collaborations based on coauthorship on publications (there are 2,837,206 publications). After analysing these publications we identified 95,100 true positive links among researchers based on co-authorship. Table 1 details true coauthorship collaborations for every research area.

Given these true relationships, we have defined three datasets. The first one, Lattes I, where true links for all eight research areas were considered, provides some general analysis. In the second and third datasets, Lattes II and Lattes III, only true links for one of the eight researcher areas were considered, allowing some specific analysis. Biological Sciences and Exact and Earth Sciences research areas were chosen, since they are the ones with more collaborations. According to cross validation principles, every dataset must be divided in training and validation sets. To avoid skewness (due to unbalanced classes), all dataset were balanced,[2] thus they have the same quantity of positive and negative examples. The positive examples were randomly chosen from the true links and the same number of negative examples were randomly collected, where negative examples means that there

[2] The problem of class skewness, imbalance in the class distribution, give rise to poor performance of a supervised learning algorithm [18]. To cope with this issue, existing research suggests several different approaches, such as altering the training sample by up-sampling or down-sampling, i.e., balancing.

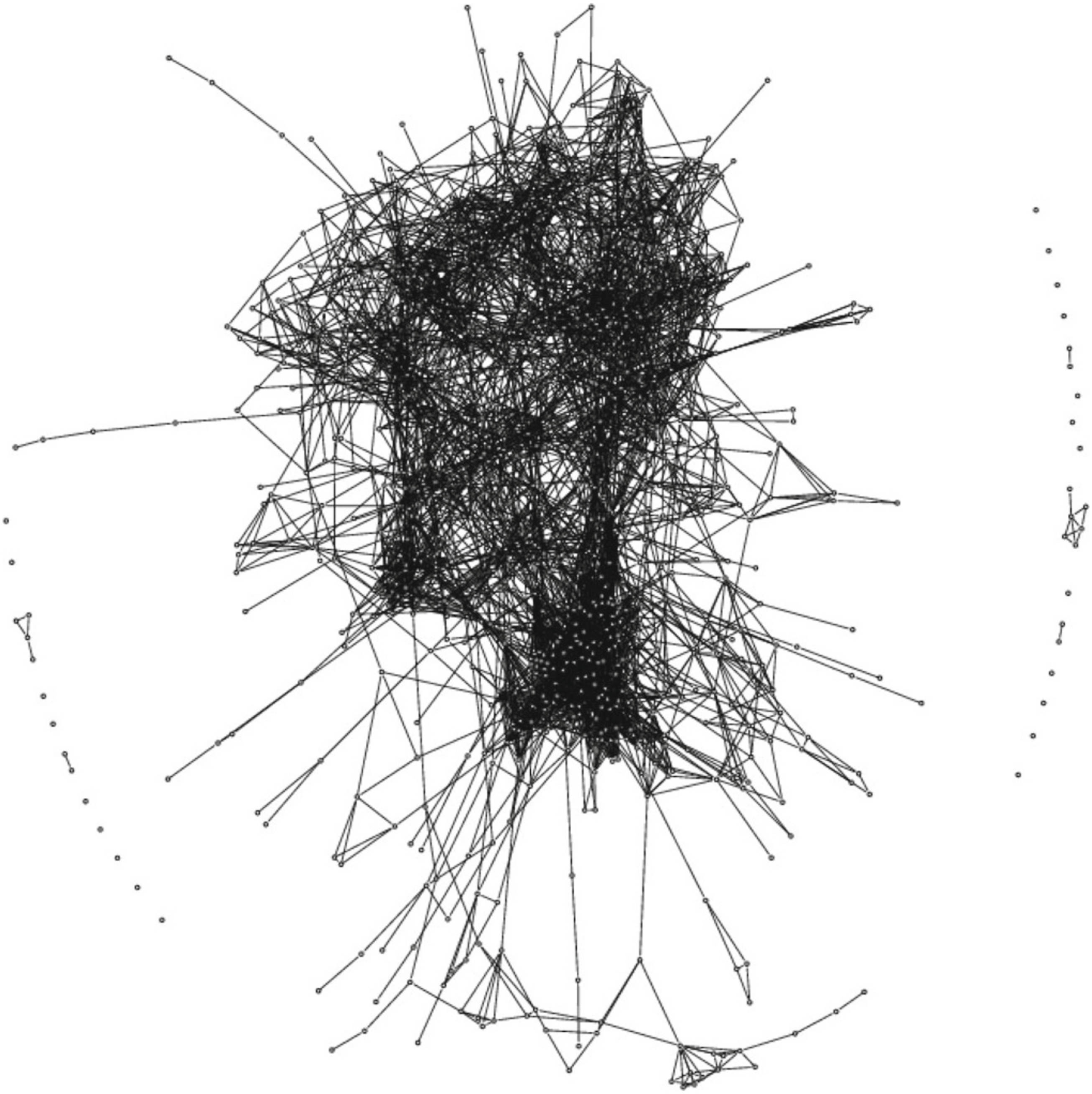

Fig. 5 Collaborations patterns in research areas (1,000 researchers): Social Sciences

is not a link between the nodes. Table 2 details the three datasets.

Although we can use probabilistic inference to decide whether there is a link between two nodes, to perform comparisons with previous approaches we resort to a classification algorithm approach. This paradigm allow us to combine several metrics (topological, semantic and probabilistic) as features of a classification algorithm. In this sense, we can compare which feature is more relevant by adding, deleting and combining features and observing the classification results.

To perform classification we resort to the Logistic regression algorithm. Which outputs values between 0 and 1 (due the logistic function) and prevent us from doing feature normalization. A threshold of 0.5 was used to decide a classification.

The features used in the classification for link prediction (defined in Sect. 2.2) are commonly extracted from topological graph properties such as neighborhood and paths between nodes. In addition, numerical features are also computed from joint probability distributions and semantics.

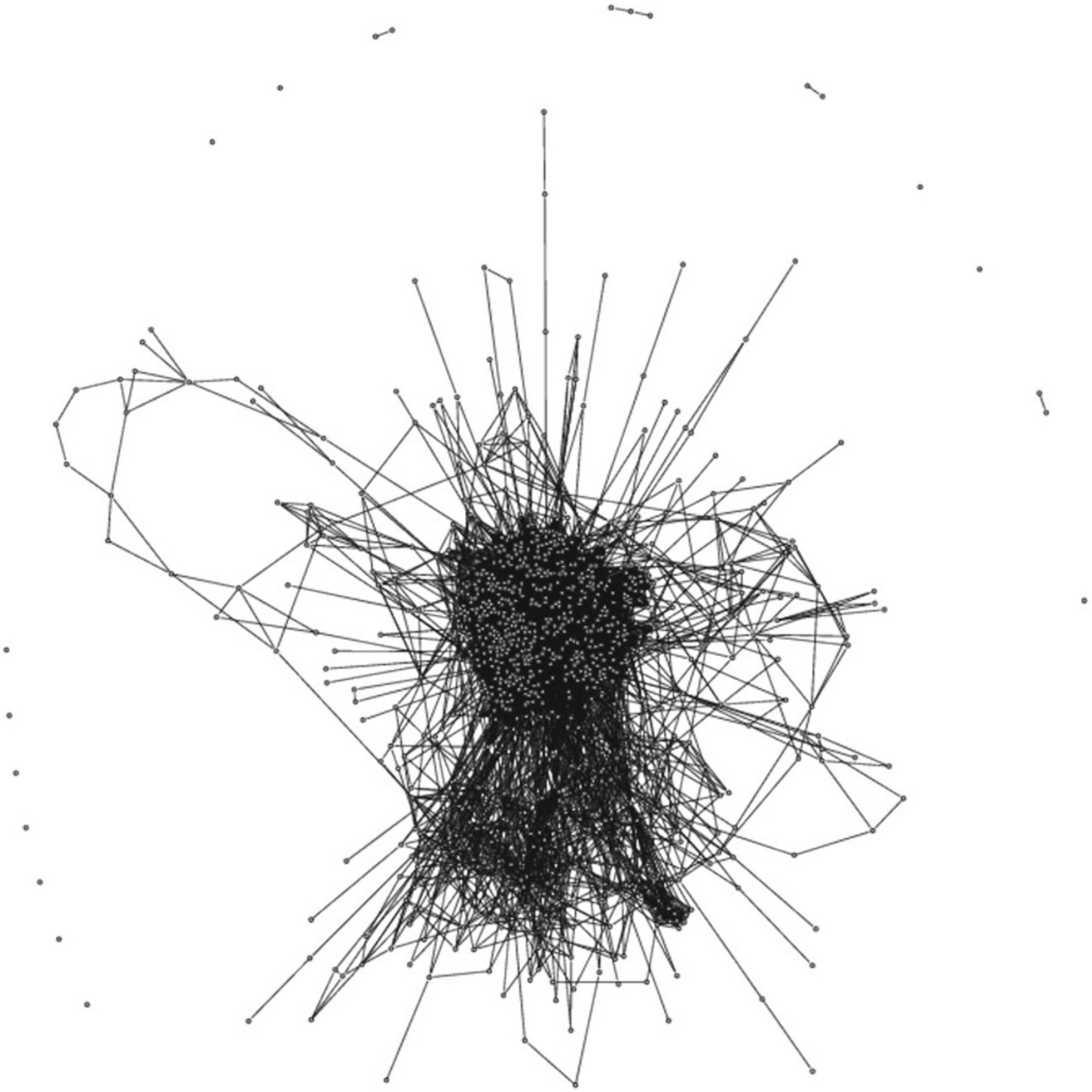

Fig. 6 Collaborations patterns in research areas (1,000 researchers): Human Sciences

The two baseline graph-based numerical feature, Katz and Adamic-Adar measures, have been used in our experiments. For the first one, since computing all paths (∞) is expensive we only consider paths of length at most four ($i \leq 4$).

We have also considered semantic features. In this work, 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 allow us to compute two features based on semantic similarity:

(i) The keyword *match* count between two researchers [10].
(ii) The *cosine* between the TFIDF features vectors of two researchers [31].

Finally, the probability $P(r(x, y)|E)$, given by our probabilistic description logic model, is also 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 coauthorships in the Lattes dataset, several experiments were run. The experiments were performed in three stages, considering incrementally, topological, semantic and probabilistic-logic scores.

In the first stage we evaluate topological scores. Two baseline scores, Katz and Adamic-Adar, have been used as fea-

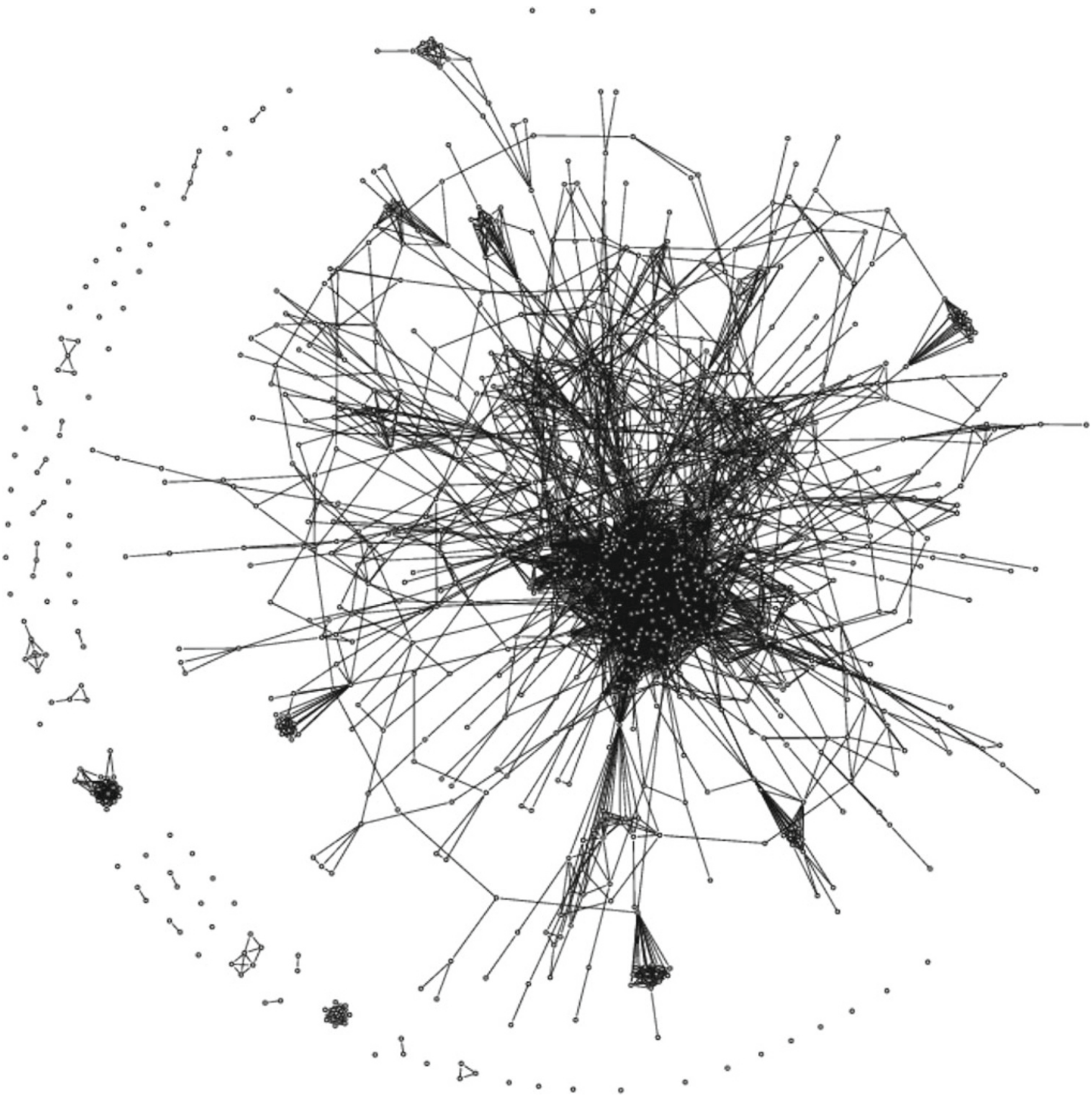

Fig. 7 Collaborations patterns in research areas (1,000 researchers): Languages and Arts

tures in the logistic regression algorithm. After a ten-fold cross validation process, the classification algorithm yielded results on accuracy which are depicted in Table 3 (stage 1).

For all three Lattes dataset, the Katz feature yields the best accuracy when the two topological features are used in isolation. Katz has been shown to be among the most effective topological measures for the link prediction task [15]. Furthermore, when we combine the two features, we improve all three accuracy.

In the second stage, we evaluate two features based on semantic similarity and their combination with topological features. Results on accuracy for these semantic features are depicted in Table 3 (stage 2). The cosine similarity feature performs better than matching keyword feature and outperforms the two former topological features. When we combine all four features together, there is an improvement in accuracy considering datasets Lattes I and Lattes II. Dataset Lattes III was indifferent to the combination of all four features.

Finally, in the third stage, a probabilistic feature based on CR$\mathcal{ALC}$ was introduced into the model. Results on accuracy for this feature are depicted in Table 3 (stage 3), showing it performs better than all other features. Moreover, there

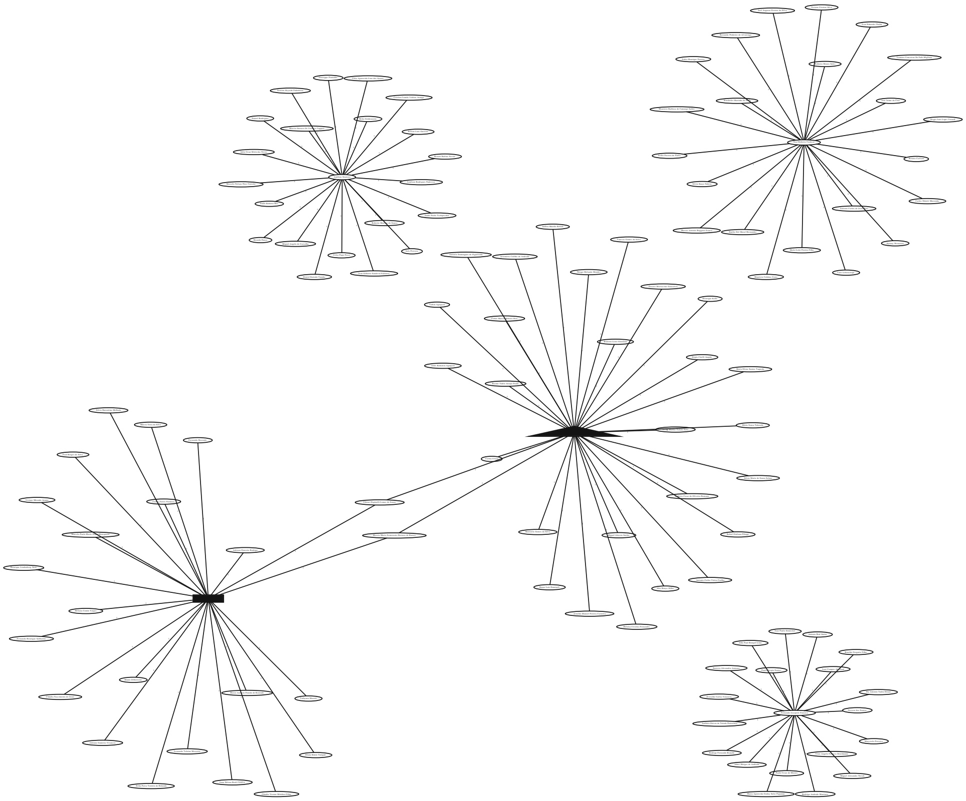

Fig. 8 Lattes collaboration network: subset of collaborations among researchers

Table 2 Lattes datasets: number of positive (+) and negative (−) examples

Name	# Examples (+/−)
Lattes I (General)	90,000
Lattes II (Biological Sciences)	20,000
Lattes III (Exact and Earth Sciences)	18,000

is significant improvement in accuracy considering datasets Lattes 1 and Lattes 2, when all five features are combined.

It is worth noting that the probabilistic logic feature used in isolation outperforms all other features and allows us to improve the classification model for link prediction on accuracy. It could be argued that such performance stems from evidence used on probabilistic inferences, but a similar analysis could be done for topological and semantic features. They use information that is missing on a probabilistic description logic setting. In conclusion, despite the fact that all features have different approaches, experimental results showed that they can be successfully used together.

Nothing prevents us from defining ad-hoc probabilistic networks to estimate link probabilities. However, by doing

Table 3 Classification results for datasets Lattes I, Lattes II and Lattes III on accuracy (%) for baseline features: Adamic-Adar (Adamic), Katz, Word matching (Match), Cosine, CrALC and a combination of them

Stage	Feature	Lattes I (acc.)	Lattes II (acc.)	Lattes III (acc.)
1	Adamic	83.34 ± 1.87	82.5 ± 1.35	81.23 ± 1.46
	Katz	85.4 ± 1.07	87.7 ± 0.91	84.43 ± 0.84
	Adamic + Katz	**85.9** ± 1.12	**87.75** ± 1.03	**85.44** ± 0.78
2	Match	75.42 ± 1.66	73.42 ± 2.66	72.8 ± 0.47
	Cosine	89.35 ± 1.28	90.4 ± 1.37	**86.7** ± 0.85
	Adamic + Katz + Match + Cosine	**91.63** ± 1.23	**90.69** ± 1.23	86.3 ± 0.12
3	Cralc	93.3 ± 0.79	94.2 ± 1.48	89.72 ± 1.67
	Adamic + Katz + Match + Cosine + Cralc	**93.89** ± 0.83	**94.46** ± 0.83	**90.2** ± 0.72

Bold values indicate the best result in the corresponding stage

Table 4 Average runtime for inference in $\textsc{cr}\mathcal{ALC}$ considering the number of nodes in the network

# Nodes	Runtime (ms)
10,000	168
100,000	175
10,000,000	185

so we are expected to define a large propositionalized network (a relational Bayesian network) [25] or estimate local probabilistic networks [31]. 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 lifted inference in $\textsc{cr}\mathcal{ALC}$ which is based on variational methods—tuned by evidence defined according to the nodes's neighborhood. Thus, for a 10,000 possible nodes, if evidence is given for 5 nodes (this is the neighborhood for a given link candidate), then there are only 6 slices which have messages interchanged. To instantiate the overall network, we use local evidence to perform inference for every link candidate, i.e., neighborhood evidence is instantiated accordingly.

In our experiments, the average runtime for inference in $\textsc{cr}\mathcal{ALC}$ (10,000 nodes network) was 168 ms. Table 4 depicts some runtime results for larger networks, which demonstrates the scalability of our approach. A direct grounding of the ontology into a propositional Bayesian network would generate an unmanageably large model.

5 Conclusion

In this paper, we have introduced a link prediction method that combines graph-based and ontological information through the use of a probabilistic description logic. Given a collaborative network, we encode interests and graph features through a $\textsc{cr}\mathcal{ALC}$ probabilistic ontology. To predict links, we resort to probabilistic inference—thus we combine and extend previous work on relational probabilistic models of link prediction, and on ontology-based link prediction. To make the proposal scalable we propose a novel strategy for approximating link probabilities: for each pair of nodes, we focus only on evidence collected along paths between them. Our proposal was evaluated on an academic domain, where links among researchers were predicted. Moreover, the approach was successfully compared with graph-based and semantic-based features.

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. Moreover, our proposal attains better scalability than previous proposals that have tried to explore probabilistic relational models for similar purposes.

Acknowledgments 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. Thanks to Jesus Pascual Mena Chalco for providing us datasets and figures of the Lattes research areas.

References

1. Adamic L, Adar E (2001) Friends and neighbors on the web. Soc Netw 25:211–230
2. Aljandal W, Bahirwani V, Caragea D, 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
3. Baader F, Nutt W (2007) Basic description logics. In: Description logic handbook. Cambridge University Press, Cambridge, pp 47–100
4. Caragea D, Bahirwani V, Aljandal W, Hsu W (2009) Ontology-based link prediction in the livejournal social network. In: SARA'09, p 1
5. Cozman FG, Polastro RB (2009) Complexity analysis and variational inference for interpretation-based probabilistic description logics. In: Proceedings of the twenty-fifth conference annual conference on uncertainty in artificial intelligence (UAI-09). AUAI Press, Corvallis, Oregon, pp 117–125
6. Fagin R, Halpern JY, Megiddo N (1990) A logic for reasoning about probabilities. Inf Comput 87:78–128
7. Getoor L, Diehl CP (2005) Link mining: a survey. ACM SIGKDD Explor Newsl 7(2):3–12
8. Getoor L, Friedman N, Koller D, Taskar B (2002) Learning probabilistic models of link structure. J Mach Learn Res 3:679–707
9. Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM (2010) A survey of statistical network models. Found Trends Mach Learn 2(2):129–233
10. Hasan MA, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In: Proceedings of SDM 06 workshop on link analysis, counterterrorism and security
11. Heinsohn J (1994) Probabilistic description logics. In: International conference on uncertainty in artificial intelligence, pp 311–318
12. Jaeger M (2002) Relational Bayesian networks: a survey. Linkoping Electr Artic Comput Inf Sci 6
13. Klinov P (2008) Pronto: A non-monotonic probabilistic description logic reasoner. In: The semantic web research and applications, pp 822–826
14. Kunegis J, Lommatzsch A (2009) Learning spectral graph transformations for link prediction. In: Proceedings of the ICML, pp 561–568
15. Liben-Nowell D, Kleinberg J (2007) The link prediction problem for social networks. J Am Soc Inf Sci Technol 7(58):1019–1031
16. Lu L, Zhou T (2011) Link prediction in complex networks: a survey. Physica A 390:1150–1170
17. Lukasiewicz T, Straccia U (2008) Managing uncertainty and vagueness in description logics for the semantic web. Semant Web J 6(4):291–308

18. Mohammad A, Mohammed J (2011) A survey of link prediction in social networks. In: Social network data analytics, pp 243–275
19. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
20. Ochoa-Luna J, Revoredo K, Cozman F (2011) Learning probabilistic description logics: a framework and algorithms. In: Proceedings of the MICAI, LNCS, vol 7094. Springer, Berlin, pp 28–39
21. Ochoa-Luna J, Revoredo K, Cozman F (2012) An experimental evaluation of a scalable probabilistic description logics approach for semantic link prediction. In: Bobillo F et al (eds) Proceedings of the 8th international workshop on uncertainty reasoning for the semantic web, vol 900. CEUR-WS.org, Shangai, China,analytics, pp 63–74
22. Ochoa-Luna J, Revoredo K, Cozman F (2012) A scalable semantic link prediction approach through probabilistic description logics. In: Proceedings of 9th artificial intelligence national meeting (ENIA)
23. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, Sananalytics, Francisco
24. Revoredo K, Ochoa-Luna J, Cozman F (2010) Learning terminologies in probabilistic description logics. In: da Rocha Costa A, Vicari R, Tonidandel F (eds) Advances in artificial intelligence SBIA, (2010) Lecture Notes in Computer Science, vol 6404. Springer/Heidelberg, Berlin, pp 41–50
25. Revoredo K, Ochoa-Luna J, Cozman F (2011) International workshop on URSW, semantic link prediction through probabilistic description logics. In: Bobillo F et al (eds) Proceedings of the 7th international workshop on URSW, vol 778, pp 87–97
26. Sachan M, Ichise R (2011) Using semantic information to improve link prediction results in network datasets. Int J Comput Theory Eng 3:71–76
27. Schmidt-Schauss M, Smolka G (1991) Attributive concept descriptions with complements. Artif Intel 48:1–26
28. Sebastiani F (1994) A probabilistic terminological logic for modelling information retrieval.In: ACM conference on research and development in information retrieval (SIGIR), pp 122–130
29. Taskar B, Wong MF, Abbeel P, Koller D (2003) Link prediction in relational data. In: Proceedings of neural information processing systems
30. Thor A, Anderson P, Raschid L, Navlakha S, Saha B, Khuller S, Zhang XN (2011) Link prediction for annotation graphs using graph summarization. In: The semantic web-ISWC, pp 714–729
31. Wang C, Satuluri V, Parthasarathy S (2007) Local probabilistic models for link prediction. In: Proceedings of the 2007 seventh IEEE ICDM. IEEE Computer Society, Washington, DC, USA, pp 322–331. doi:10.1109/ICDM.2007.108
32. Wohlfarth T, Ichise R (2008) Semantic and event-based approach for link prediction. In: Proceedings of the 7th international conference on practical aspects of knowledge management