Semantic Link Prediction through Probabilistic Description Logics

Kate Revoredo
Department of Applied Informatics

José Eduardo Ochoa Luna and Fabio Cozman
Escola Politécnica

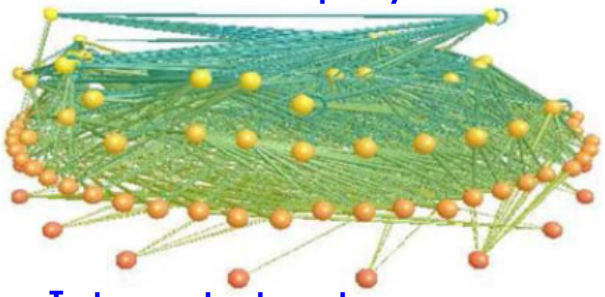
Outline

- **Introduction**
- Background knowledge
- Proposal: Link Prediction using CrALC
- Preliminary Results
- Conclusion and perspective

Introduction

A network can describe social, biological, information systems ....

Predator - prey



Paris subway

Internet structure

Research collaboration

- In a network
  - Nodes represent objects, individuals
  - Links denote relations or interactions between the nodes

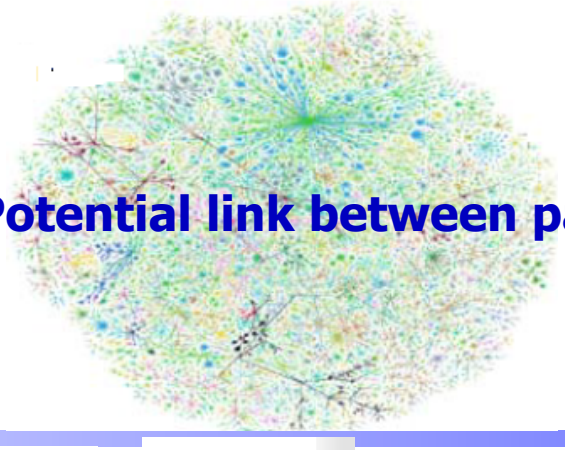
Introduction

Automatic prediction of possible links in a network is an interesting issue.

Predator - prey

Paris subway

Internet structure



Research collaboration

Introduction

- **Link prediction** aims at predicting whether two nodes should be connected given that previous informations about their relationships or interests are known.
- Possibilities
  - Network structure analysis
    - Numerical informations about the nodes are analyzed
  - Object knowledge analysis
    - Semantic related to the domain of the objects are considered
  - A combination of them

Introduction

- Knowledge about the domain can be formalize using **ontology**.
  - Description logic (DL) can be the language used by the ontology

Introduction

- DL for the Academic domain....

$$Researcher \equiv Person \sqcap \exists hasPublication.Publication$$

$$Student \equiv Person \sqcap \exists hasAdvise.Researcher$$

$$Collaborator \equiv Researcher \sqcap \exists sharePublication.Researcher$$

$$Researcher \sqsubseteq Professor$$

- And if there is uncertainty about the domain?
  - Not all researcher is a professor

Introduction

- Uncertainty about the domain can be formalize using **probabilistic ontology**.
  - Probabilistic Description logic (PDL) can be the language used by the probabilistic ontology
    - P-Classic [KOLLER et.al.,97]
    - P-SHOIN [Lukasiewicz,07]
    - PR-OWL [Costa et.al.,06]
    - **CrALC logic [Polastro et.al.,08]**

Proposal

- How to predict a new link in a network considering knowledge about the domain and the uncertainty involved?
  - Using an algorithm for link prediction that considers semantic and uncertainty about the domain through the use of the PDL CrALC.

Outline

- Introduction
- **Background knowledge**
  - **Probabilistic Description Logic CrALC**
- Proposal: Link Prediction using CrALC
- Preliminary Results
- Conclusion and perspectives

Probabilistic description logic CrALC

- CrALC
  - Is a probabilistic extension of the DL ALC
    - Keep all constructors
    - Add probabilistic inclusions such as
      - P(Researcher | Person) = α
      - Semantic: $\forall x \in D \mid P(Researcher(x) \mid Person(x)) = \alpha$
  - Adopts an interpretation-based semantics

Learning crALC

- A PDL crALC can be learned automatically from data [Revoredo, et.al., 2010].

Inference in CrALC

- CrALC assumes an acyclic terminology (T), thus T can be represented through a directed acyclic graph g(T)
  - Each concept name and role name is a node in g(T)
  - If a concept C direclty uses concept D, then D is a parent of C in g(T)
  - Each existencial restriction (∃r.C) and value restriction (∀r.C) is added to the graph g(T) as nodes
    - An edge from role r to each restriction directly using it is added
    - Each restriction node is a deterministic node
  - Relational Bayesian Network (RNB) [Jeager,02]
- Probabilistic inference is computed in the propositionalization of the graph.
  - Exact and approximate algorithms

Inference in CrALC - Example

$B \sqsubseteq A$

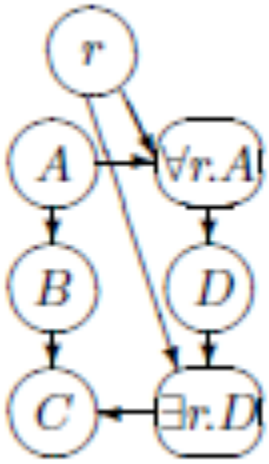
$C \sqsubseteq B \sqcup \exists r.D$

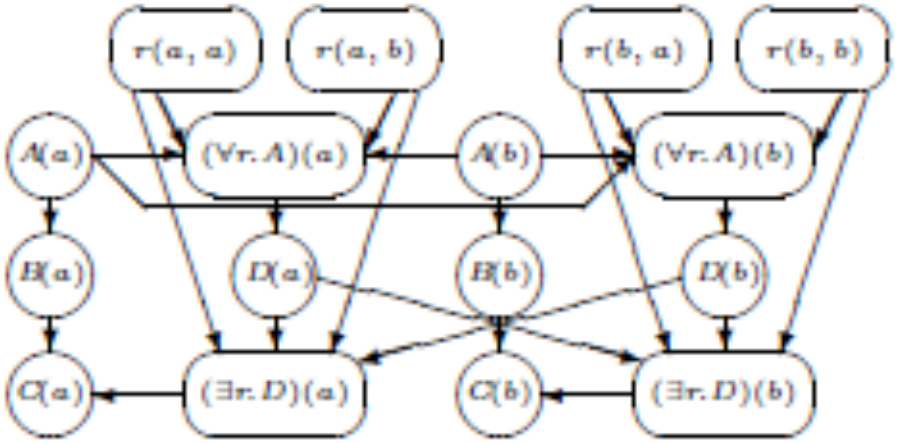
$P(A)=0.9,$

$P(B|A)=0.4$

$P(C \mid B \sqcup \exists r.D)=0.6$

$P(D|\forall r.A)=0.3$





- $P(D(a)|B(b)) = 0.232$

Outline

- Introduction
- Background knowledge
- **Proposal: Link Prediction using CrALC**
- Preliminary Results
- Conclusion and perspective

Example

- In a collaboration network
  - Objects: researchers
  - Relationship: "share a publication"

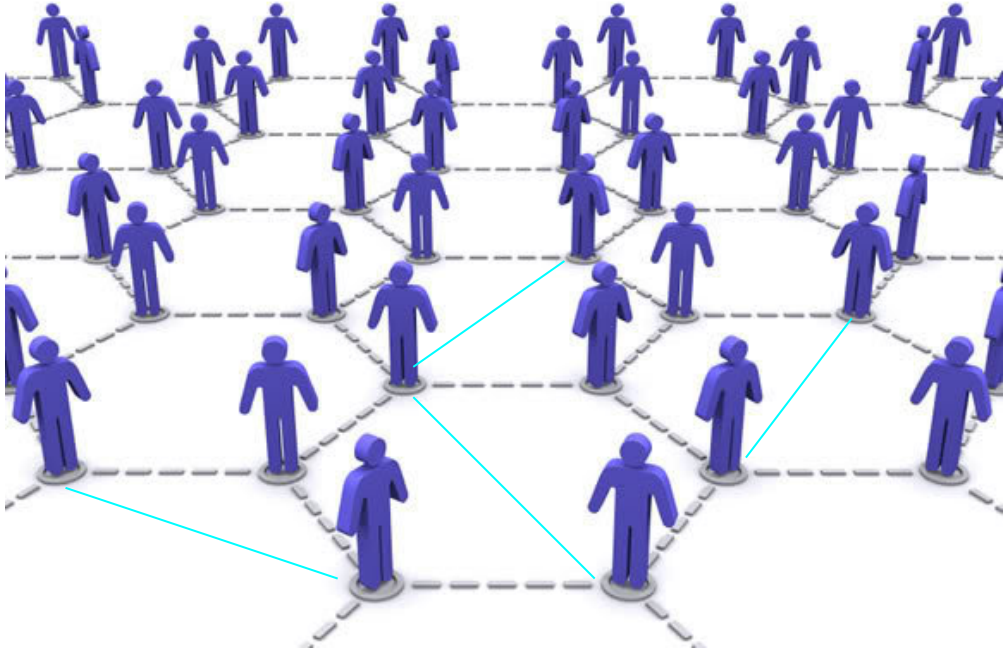
- PDL crALC describing the domain
  - Concepts:
    - Researcher
    - P(Publication)=0.3
    - P(NearCollaborator | Researcher ⊓ ∃sharePublication. ∃hasSameInstitution. ∃sharePublication.Researcher) = 0.95
    - StrongRelatedResearcher ≡ Researcher ⊓ (∃sharePublication.Researcher ⊓ ∃wasAdvised.Researcher)

      ⋮
  - Roles
    - hasPublication
    - P(sharePublication)=0.22
    - P(hasSameInstitution)=0.14

Link Prediction using CrALC - Task

- Given
  - A network **N** defining relationships between objects;
  - An ontology **O,** represented by crALC, describing the domain;
  - The ontology role **r** that defines the semantic of the relationship between objects in the network;
  - The ontology concept **C** that describes the network objects.
- Find
  - A revised network $\mathbf{N}_f$ with new relationships between objects.

Proposal - Example



- Since the links correpond to a role in the PDL crALC, a new link is added if the probability of the role for the respectively objects given some evidence is high
  - P(sharePublicaton(ann,mark)|evidence)=0.87

Algorithm

- **Require**: network *N*, ontology *O*, role *r(_,_)*, concept *C*, *threshold*
- **Ensure**: network N_f
  - Define N_f as *N*
  - **For** all pair of instances *(a,b)* of concept *C* **do**
    - **If** does not exist a link between nodes *a* and *b* in the network *N* **then**
      - Infer probability *P(r(a,b)|evidences)* using the RBN created through the ontology *O*
      - **If** *P(r(a,b)|evidences) > threshold* **then**
        - » Add a link between *a* and *b* in the network N_f

- Alternatively to the threshold, the top-k infered links, where k would be a parameter, can be included.

Outline

- Introduction
- Background knowledge
- Proposal: Link Prediction using CrALC
- **Preliminary Results**
- Conclusion and perspective

Preliminary Results

- Collaboration network of researchers
- Data gathered from Lattes Curriculum Platform
  - Public repository of Brazilian researcher curriculum
  - Informations: name, address, education, professional experience, areas of expertise, publication ....
  - 1200 researches randomly selected and structured as

$$\text{Researcher}(r1), \text{Researcher}(r2), \text{Researcher}(r4), \ldots$$
$$\text{wasAdvised}(r8, r179), \text{wasAdvised}(r30, r83), \text{wasAdvised}(r33, r1), \ldots$$
$$\text{sharePublication}(r1, r32), \text{sharePublication}(r4, r12), \text{sharePublication}(r5, r115), \ldots$$
$$\text{sameExaminationBoard}(r1, r32), \text{sameExaminationBoard}(r4, r12), \ldots$$
$$\text{hasSameInstitution}(r1, r27), \text{hasSameInstitution}(r1, r28), \ldots$$
$$\text{advises}(r1, r33), \text{advises}(r1, r171), \text{advises}(r1, r81), \ldots$$

Preliminary Results

- Using the data, a PDL crALC was learned [Revoredo et,al., 2010]

$P(\text{Researcher}) = 1.0$ $P(\text{wasAdvised}) = 0.29$
$P(\text{hasSameInstitution}) = 0.83$ $P(\text{sharePublication}) = 0.73$
$P(\text{sameExaminationBoard}) = 0.41$

$P(\text{NearCollaborator} \mid \text{Researcher} \sqcap \exists \text{sharePublication}.\exists \text{hasSameInstitution}.\exists \text{sharePublication}.\text{Researcher}) = 0.95$

$\text{FacultyNearCollaborator} \equiv \text{NearCollaborator} \sqcap \exists \text{sameExaminationBoard}.\text{Researcher}$

$P(\text{NullMobilityResearcher} \mid \text{Researcher} \sqcap \exists \text{wasAdvised}.\exists \text{hasSameInstitution}.\text{Researcher}) = 0.98$

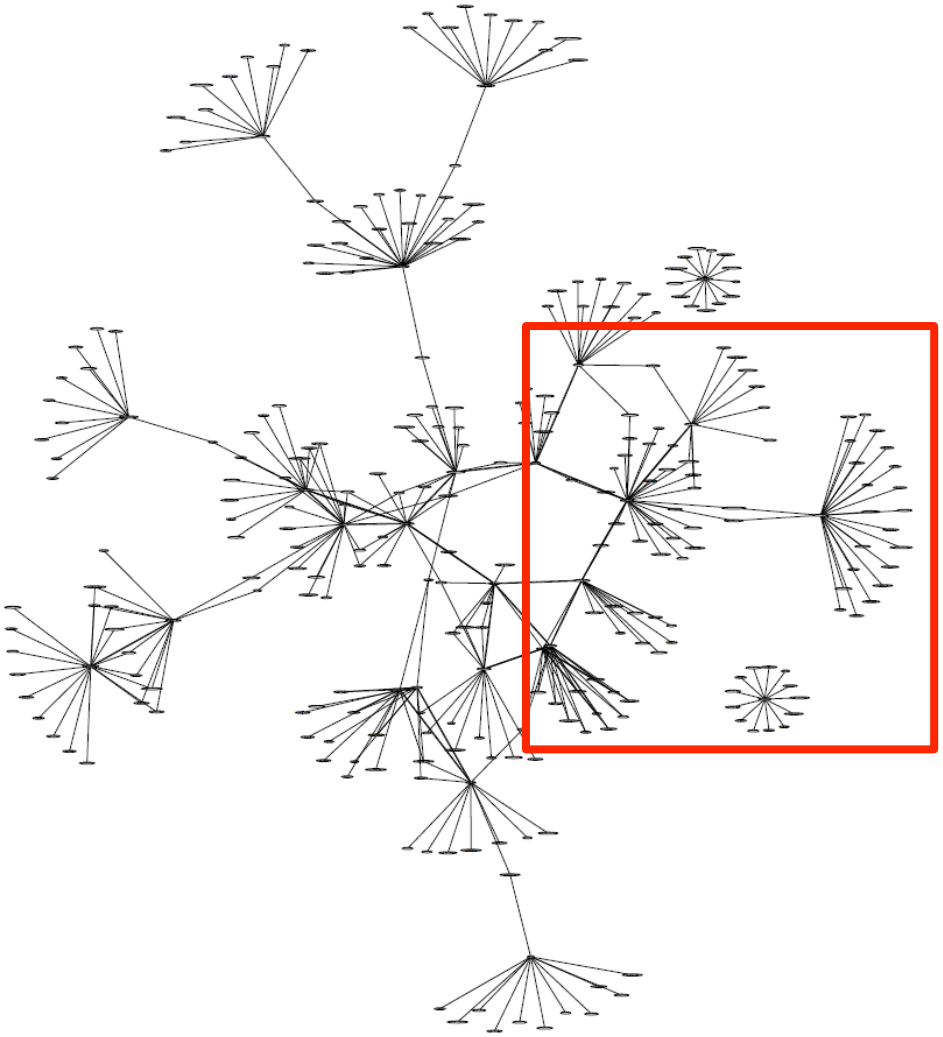
$\text{StrongRelatedResearcher} \equiv \text{Researcher} \sqcap (\exists \text{sharePublication}.\text{Researcher} \sqcap \exists \text{wasAdvised}.\text{Researcher})$

$\text{InheritedResearcher} \equiv \text{Researcher} \sqcap (\exists \text{sameExaminationBoard}.\text{Researcher} \sqcap \exists \text{wasAdvised}.\text{Researcher})$

- Object: instances of concept Researcher
- Relationships: role sharePublication

Preliminary Results

- Using the data, a collaboration network was learned
  - Object: instances of concept Researcher
  - Relationships: role sharePublication
  - 303 researchers that share a publication were found

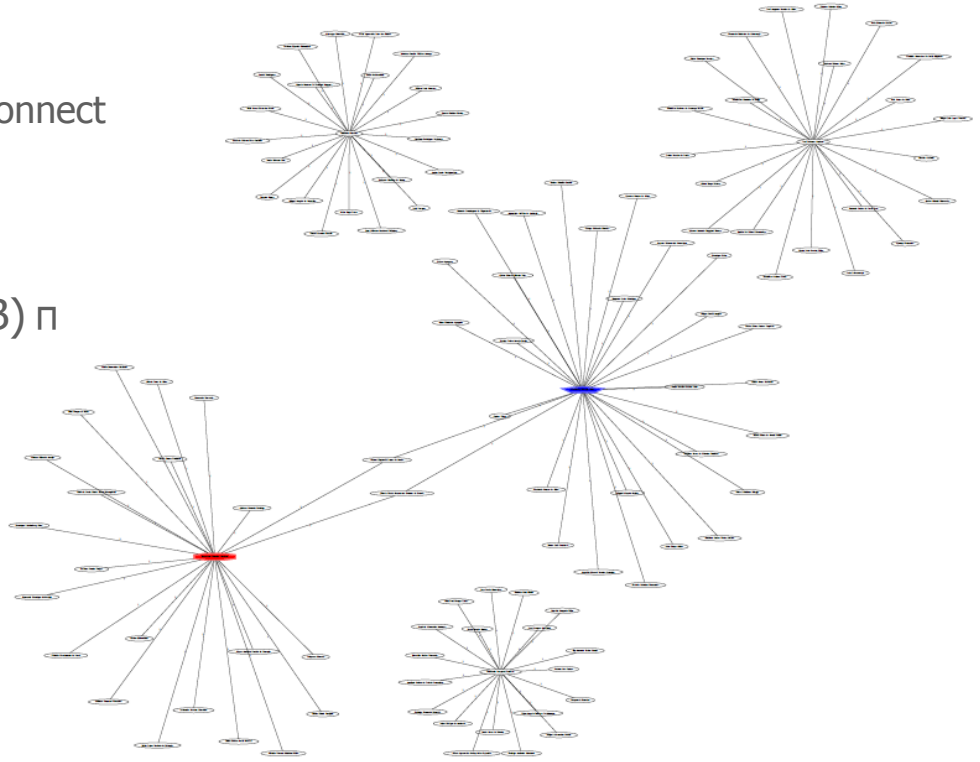


- The proposal algorithms were run and some links were proposed
- Moreover...

Preliminary Results

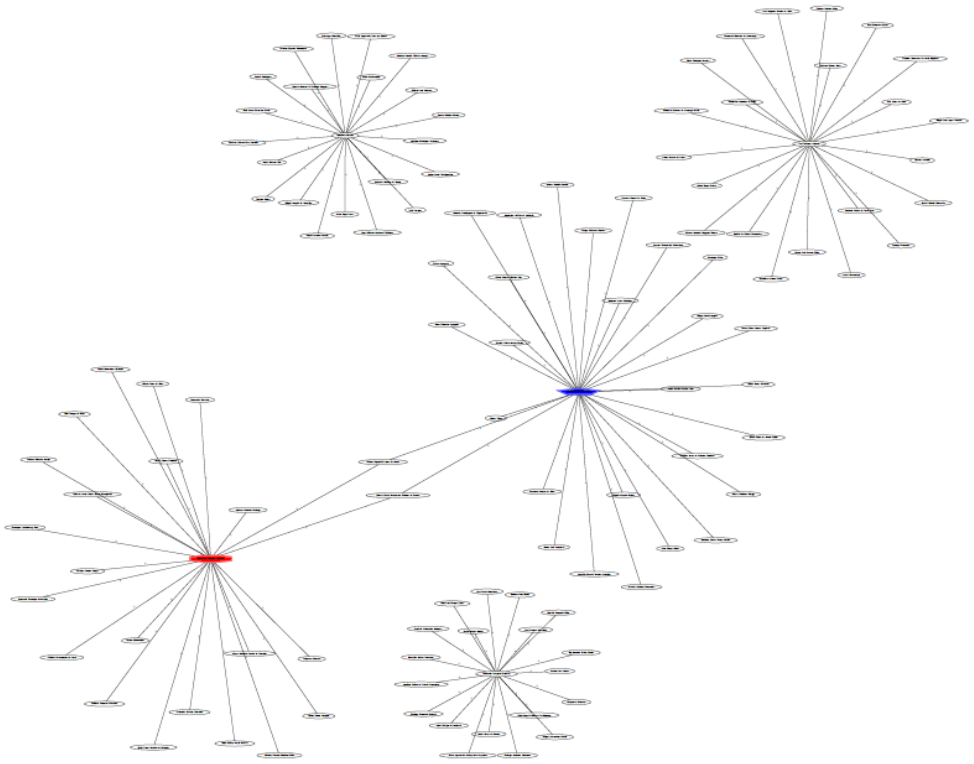
- A more guided link prediction: Links among researchers from different groups
  - Infer P(link(Red,Blue)|evidence)
  - P(PublicationCollaborator(R)|Researcher(R) ⊓ ∃hasSameInstitution.Researcher(B))=0.57

- more evidence was gained...
  - Information about nodes that indirectly connect these 2 groups (I1,I2)
  - P(PublicationCollaborato(R)| Researcher(R) ⊓ ∃hasSameInstitution.Researcher(B) ⊓ ∃sharePublication(I1). ∃sharePublication(B) ⊓ ∃sharePublicaton(I2). ∃sharePublication(B))=0.65



Preliminary Results

- A more guided link prediction: Links among researchers in the same group
  - For each i=1,...,k and j=1,...,n
    - Infer P(link(Red_i,Red_j)|evidence) e P(link($Blue_i$,$Blue_j$)|evidence)



Conclusion

- An approach for predicting links in a network using the probabilistic description logic CrALC was proposed
  - In the network
    - Objects represents instances of a concept in the PDL crALC
    - Links represents a role in the PDL crALC
  - Inference with the PDL crALC indicates links that should be included in the network
- Experiments with Lattes Curriculum Plataform showed the potential of the idea.

Perspectives

- Consideration of probabilistic networks
  - Since the new links came from probabilistic inference, a weight in the link can be considered
- Applications to larger domains

Acknowledgements

- CAPES
- CNPq
- FAPESP – projeto 2008/03995-5

Thank you!