# An Experimental Evaluation of a Scalable Probabilistic Description Logic Approach for Semantic Link Prediction

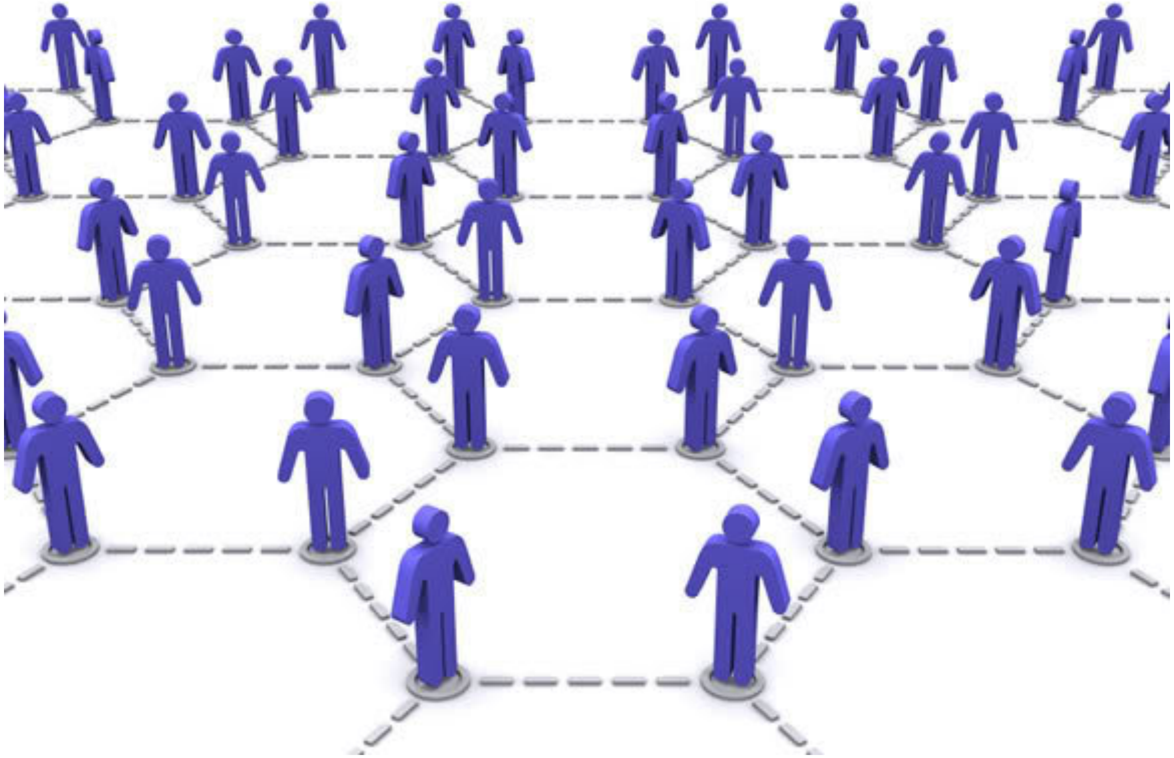
**Kate Revoredo**
**Department of Applied Informatics**

UNIRIO

**José Eduardo Ochoa Luna and Fabio Cozman**
**Escola Politécnica**

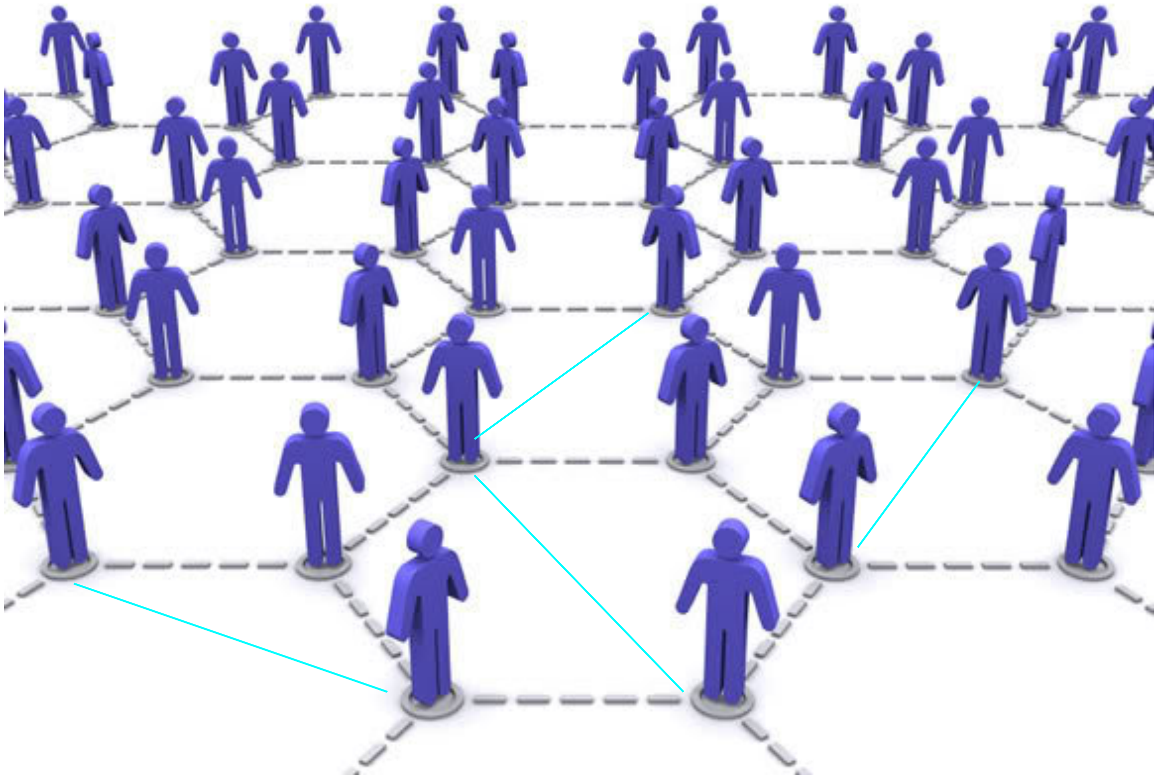
USP

# Motivation



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

# Motivation



- How to predict automatically a link?

# Motivation

- Possibilities of **link prediction**
  - 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

- There is **uncertainty** about the predicted link.

UNIRIO

# Problem

- How to predict a link in a network considering knowledge about the domain, the uncertainty involved and in a scalable way?

# Introduction

- Knowledge about the domain can be formalize using **Ontology**.
  - **Description logic** is a language used to represent Ontology
    - 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

UNIRIO USP

# Introduction

- Uncertainty about the domain can be formalize using **probabilistic ontology**.
  - **Probabilistic Description Logic** is a language used to represent probabilistic ontology
    - P-Classic [KOLLER et.al.,97]
    - P-SHOIN [Lukasiewicz,07]
    - PR-OWL [ Costa et.al.,06]
    - **Credal $\mathcal{ALC}$ (Cr$\mathcal{ALC}$) logic [Polastro et.al.,08]**

UNIRIO

# Proposal

- An algorithm for link prediction that through probabilistic description logic Cr$\mathcal{ALC}$
  - considers domain semantic
  - considers domain uncertainty
  - it is scalable.

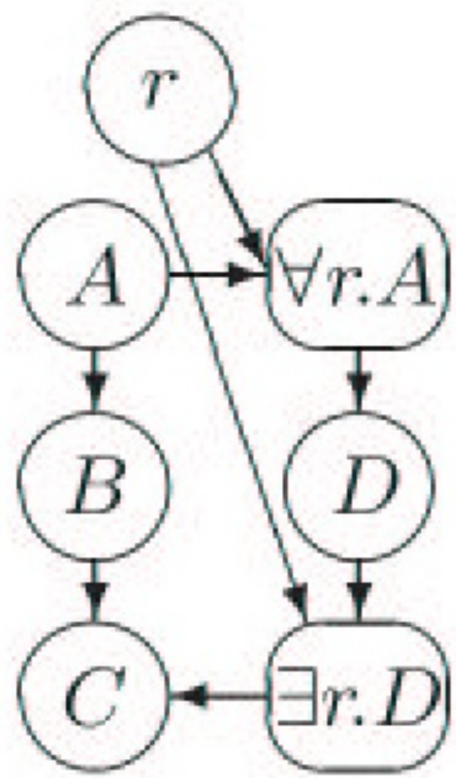
UNIRIO

# Outline

- Introduction
- **Probabilistic Description Logic Cr$\mathcal{ALC}$**
- Link Prediction using Cr$\mathcal{ALC}$
- Experimental Results
- Conclusion and perspectives

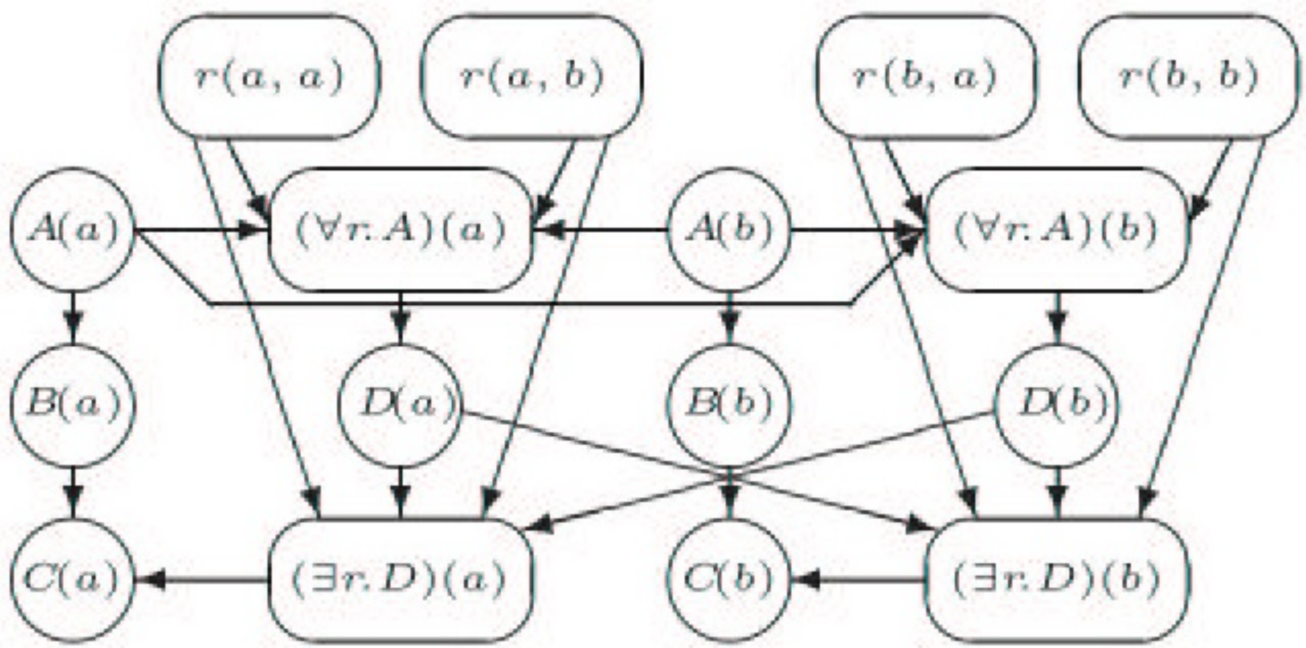
# Cr$\mathcal{ALC}$ - Example

$B \sqsubseteq A$
$C \sqsubseteq B \sqcup \exists r.D$
$P(A)=0.9,$
$P(B|A)=0.4$
$P(C \mid B \sqcup \exists r.D)=0.6$
$P(D|\forall r.A)=0.3$



# Inference in Cr$\mathcal{ALC}$ - Example

- *Domain*={a,b}




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

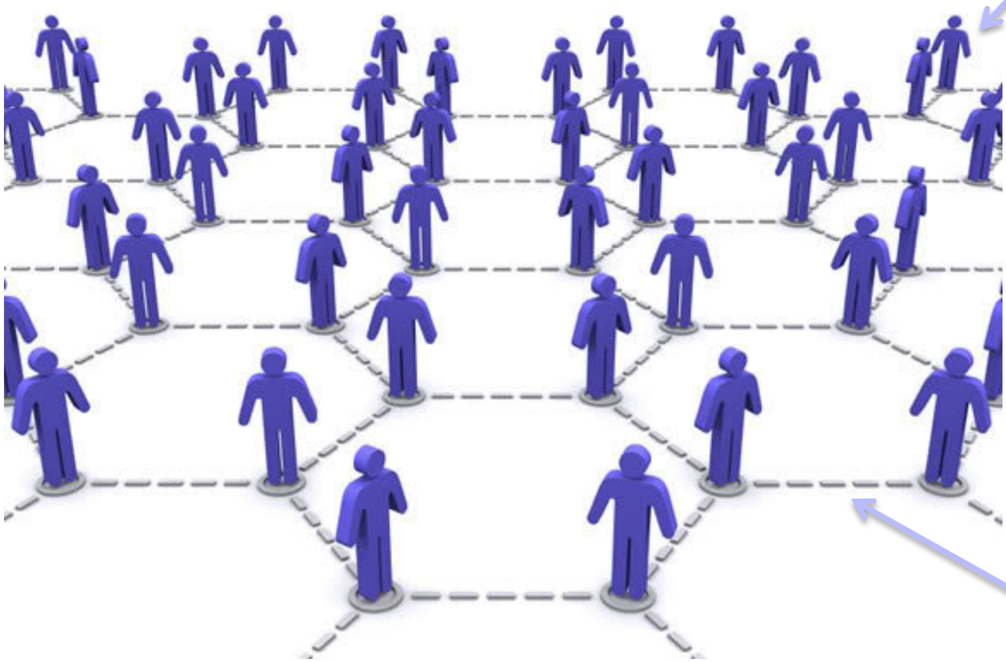
# Outline

- Introduction
- Probabilistic Description Logic Cr$\mathcal{ALC}$
- **Link Prediction using CrALC**
- Experimental Results
- Conclusion and perspective

UNIRIO USP

# Example

- In a collaboration network
  - Objects: researchers
  - Relationship: “share a publication”

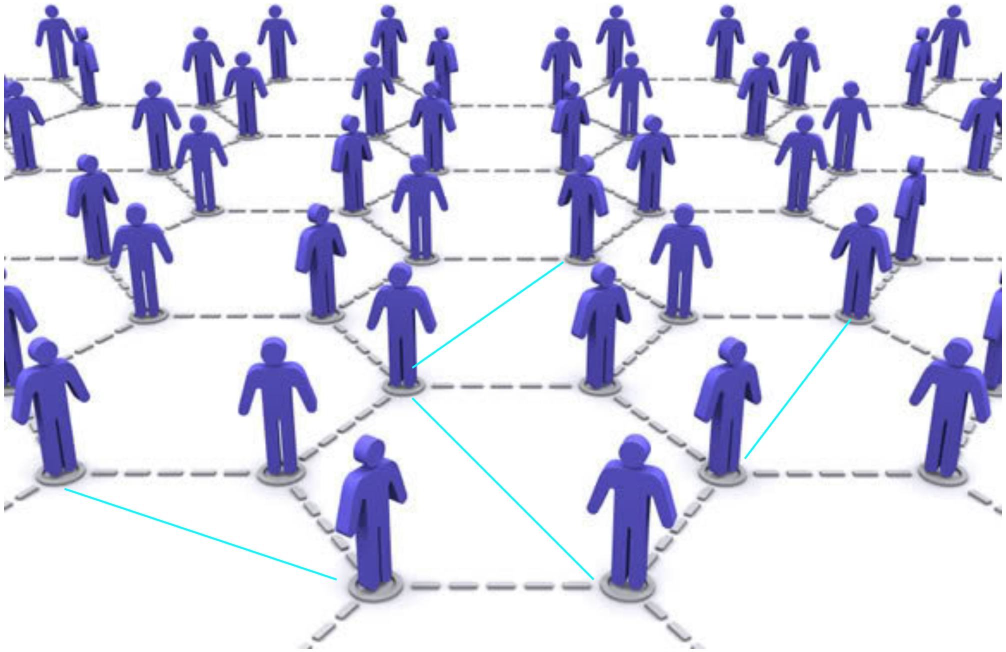


- cr$\mathcal{ALC}$ describing the domain
  - Concepts:
    - Researcher
    - P(Publication)=0.3
    - P(NearCollaborator | Researcher $\sqcap$ $\exists$sharePublication. $\exists$hasSameInstitution. $\exists$sharePublication.Researcher) = 0.95
    - StrongRelatedResearcher $\equiv$ Researcher $\sqcap$ ($\exists$sharePublication.Researcher $\sqcap$ $\exists$wasAdvised.Researcher)

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

UNIRIO USP

# Proposal - Example



- Since the links correpond to a role in cr$\mathcal{ALC}$, a new link is added if the probability of the role for the respectively objects given some evidence is high
  - P(sharePublication(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.

# Algorithm

- For every individual of the domain a "slice" in the RBN is considered
  - *All slices without evidence are consolidated in one [Cozman and Polatro, 2009] to optimize inference algorithm.*
- **Less individuals with evidence → faster inference**

- In social networks many individuals are considered
  - Usually there is evidence for each one.
- The algorithm proposed may not scale.
- We need an approximation.
  - When computing P(r(a,b)|evidences) only evidences of a, b and the individuals **most related** to them should be considered.
    - Graph-based features are considered

# Outline

- Introduction
- Probabilistic Description Logic CrALC
- Link Prediction using CrALC
- **Experimental Results**
  - **Scenario**
  - **Methodology**
  - **Results**
- Conclusion and perspective

# Scenario

- 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 ....
  - 1100 researches randomly selected and structured as

Researcher(r1), Researcher(r2), Researcher(r4), . . .
wasAdvised(r8, r179), wasAdvised(r30, r83), wasAdvised(r33, r1), . . .
sharePublication(r1, r32), sharePublication(r4, r12), sharePublication(r5, r115), . . .
sameExaminationBoard(r1, r32), sameExaminationBoard(r4, r12), . . .
hasSameInstitution(r1, r27), hasSameInstitution(r1, r28), . . .
advises(r1, r33), advises(r1, r171), advises(r1, r81), . . .

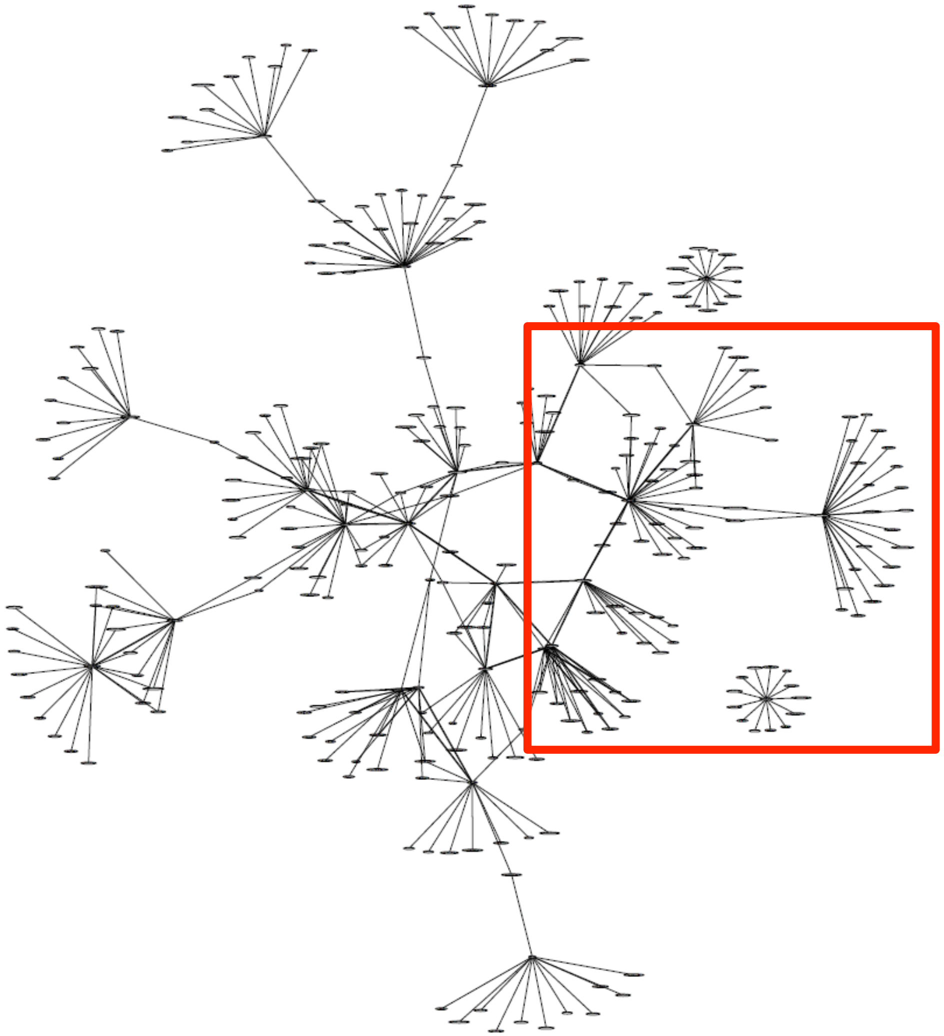
# Scenario

- Using the data, a cr$\mathcal{ALC}$ was learned [Revoredo et,al., 2010]

$$
\begin{array}{ll}
P(\mathsf{Researcher}) = 1.0 & P(\mathsf{wasAdvised}) = 0.29 \\
P(\mathsf{hasSameInstitution}) = 0.83 & P(\mathsf{sharePublication}) = 0.73 \\
P(\mathsf{sameExaminationBoard}) = 0.41 & \\
P(\mathsf{NearCollaborator} & \mid \mathsf{Researcher} \sqcap \exists \mathsf{sharePublication}.\exists \mathsf{hasSameInstitution}. \\
 & \exists \mathsf{sharePublication}.\mathsf{Researcher}) = 0.95 \\
\mathsf{FacultyNearCollaborator} \equiv & \mathsf{NearCollaborator} \\
 & \sqcap\ \exists \mathsf{sameExaminationBoard}.\mathsf{Researcher} \\
P(\mathsf{NullMobilityResearcher} & \mid \mathsf{Researcher} \sqcap \exists \mathsf{wasAdvised}. \\
 & \exists \mathsf{hasSameInstitution}.\mathsf{Researcher}) = 0.98 \\
\mathsf{StrongRelatedResearcher} \equiv & \mathsf{Researcher} \\
 & \sqcap\ (\exists \mathsf{sharePublication}.\mathsf{Researcher} \sqcap \\
 & \exists \mathsf{wasAdvised}.\mathsf{Researcher}) \\
\mathsf{InheritedResearcher} \equiv & \mathsf{Researcher} \\
 & \sqcap\ (\exists \mathsf{sameExaminationBoard}.\mathsf{Researcher} \sqcap \\
 & \exists \mathsf{wasAdvised}.\mathsf{Researcher})
\end{array}
$$

# Scenario

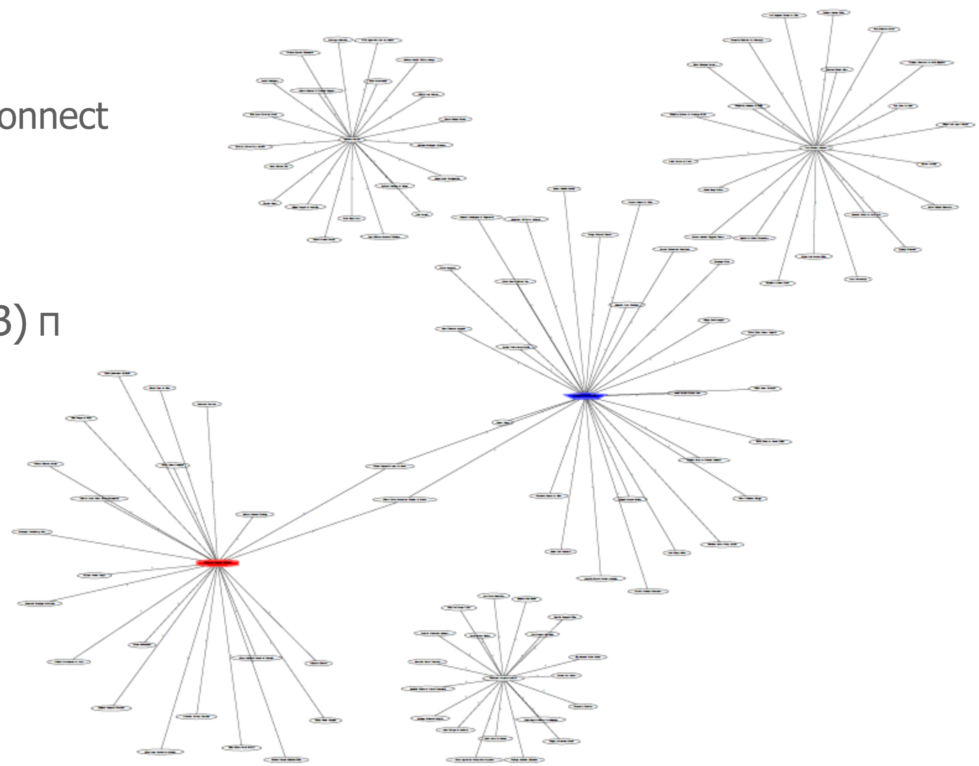
- Using the data, a collaboration network was learned
  - Object: instances of concept Researcher
  - Relationships: role sharePublication
  - 303 researchers that share a publication were found



- .

# Scenario

- A more guided link prediction: Links among researchers from different groups
  - Infer P(link(Red,Blue)|evidence)
  - $P(PublicationCollaborator(R)|Researcher(R) \sqcap \exists hasSameInstitution.Researcher(B))=0.57$
- more evidence was gained...
  - Information about nodes that indirectly connect these 2 groups (I1,I2)
  - $P(PublicationCollaborato(R)| Researcher(R) \sqcap \exists hasSameInstitution.Researcher(B) \sqcap \exists sharePublication(I1). \exists sharePublication(B) \sqcap \exists sharePublicaton(I2). \exists sharePublication(B))=0.65$



# Methodology

- A logistic-regression classification algorithm was trained and used to evaluate our proposal.
  - Features are graph-based characteristics (neighbors nodes, path between nodes,…)
  - For comparison:
    - Structural characteristics
      - Katz measure [Liben-Nowell and Kleinberg 2003]
        - » weighted sum of the number of paths in the graph that connect two nodes
          - » higher weight for shorter paths
          - » Paths of length at most 4
      - Adamic-Adar measure [AdamicandAdar2001]
        - » which computes the similarity between two nodes in a graph
          - » weight the hub nodes less and rarer nodes more
    - Semantic characteristics: a researcher is represented by the of words appearing in the title of its publications (stop words removed)
      - keyword match count between two researchers
      - cosine similarity applied between two researchers (vector representation with TFIDF)

# Methodology

- Dataset with information about 1100 researchers
  - Positive instances: there is a link between two researchers
  - Negative instances: there is **not** a link between two researchers.
- 10-fold cross-validation

# Results

| | Adamic | Katz | Adamic+Katz |
|---|---|---|---|
| Accuracy | 72,25 ±1.87 | 75.49 ± 2.07 | 76.44 ± 2.03 |

| | Match | Cosine | Adamic+Katz+Match+Cosine |
|---|---|---|---|
| Accuracy | 69.42 ±2.66 | 82.45 ±1.37 | 85.63 ± 1.23 |

| | CrALC | Adamic+Katz+Match+Cosine+CrALC |
|---|---|---|
| Accuracy | 87.72 ±0.52 | 89.48 ± 0.96 |

- Time for inference: 43.401 miliseconds
- Computation not possible without the approximation proposal

# Conclusion

- An approach for predicting links in a network using the probabilistic description logic Cr$\mathcal{ALC}$ was proposed
  - In the network
    - Objects represents instances of a concept in cr$\mathcal{ALC}$
    - Links represents a role in cr$\mathcal{ALC}$
  - Inference with cr$\mathcal{ALC}$ indicates links that should be included in the network
- Experiments with Lattes Curriculum Plataform showed the potential of the idea.
- Do to the approximation proposal the approach scales

# Perspectives

- Other metrics to reduce the number of evidences considered during inference
- 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


- CNPq
- FAPESP – projeto 2008/03995-5
- FAPERJ grant E-26/111484/2010

# Questions?