Link Prediction in Large Directed Graphs

Dario Garcia-Gasulla
Ulises Cortés

Overview

- ◎Motivation
- ◎State of the Art
- ◎Hypothesis
- ◎Hierarchical Link Prediction
- ◎Computational Models
- ◎Data Sets & Results
- ◎Conclusions
- ◎Discussion & Future Work

Overview

- ◎Motivation
- ◎State of the Art
- ◎Hypothesis
- ◎Hierarchical Link Prediction
- ◎Computational Models & Designs
- ◎Data Sets & Results
- ◎Conclusions
- ◎Discussion & Future Work

Motivation

Objects Data **INTERNET!** *Object-object* Data

In Data Mining and Machine Learning …

From *intra-entity* to *inter-entity* patterns

"One small step for data, one giant leap for data science"

Motivation

New family of domains

-Web graphs
-Social networks
-Biological networks
-Product recommendation
-Terrorist associations
-...

Typically LARGE

- but, *how large*?

Motivation

Whole new set of problems

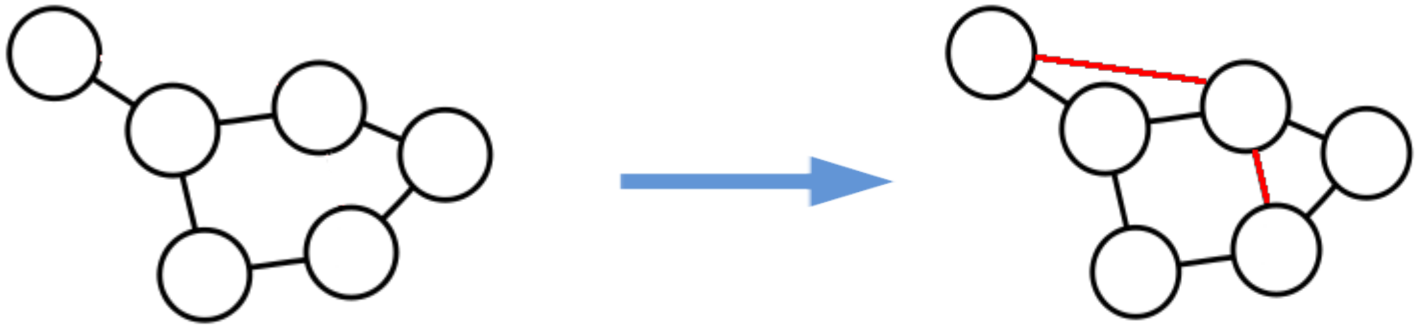
- Rank entities based on importance
- Find groups of entities
- Discover association patterns
- Predict new relations

Let us call it just **Graph Mining**

Motivation

Link Prediction

- Find new relations given the structure of a graph



Motivation

Link Prediction

Needle in a haystack

-How many friends you do have in Facebook?

-How many friends you do NOT have?

we need **PRECISION**

An ocean of variables depending on one another

Friends define friendship

we need **SCALABILITY**

Overview

- ◎Motivation
- ◎State of the Art
- ◎Hypothesis
- ◎Hierarchical Link Prediction
- ◎Computational Models & Designs
- ◎Data Sets & Results
- ◎Conclusions
- ◎Discussion & Future Work

State of the Art

Compute statistics on the **graph**

- Bayes / Markov (Getoor and Taskar, 2007)
- Tensors (Nickel et al., 2011)

Compute the likelihood of the **graph**

- Hierarchies (Clauset et al., 2008)
- Communities (Stochastic block models)

Compute **entity-entity** similarities

- Number of paths

State of the Art

Similarity-based Link Prediction

-Scalable 
-Parallelizable 
-Unprecise 

We look for common neighbors... *how far*?

-*Local*: 2-steps. It works, but not well enough.
-*Global*: No limit. Poor scaling. Disappointing results.
-*Quasi-local*: Unknown variable distance. Best!
But wait, *unknown* distance?

State of the Art

Similarity-based: The essence

-How many common neighbors we have? (Newman, 2001)

-How many rare common neighbors we have? (Adamic and Adar, 2003) (Zhou, 2009)

Common Neighbors	$s_{x,y}^{CN} = \|\Gamma(x) \cap \Gamma(y)\|$
Adamic/Adar	$s_{x,y}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(\|\Gamma(z)\|)}$
Resource Allocation	$s_{x,y}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\|\Gamma(z)\|}$

Overview

- Motivation
- State of the Art
- Hypothesis
- Hierarchical Link Prediction
- Computational Models & Designs
- Data Sets & Results
- Conclusions
- Discussion & Future Work

Hypothesis

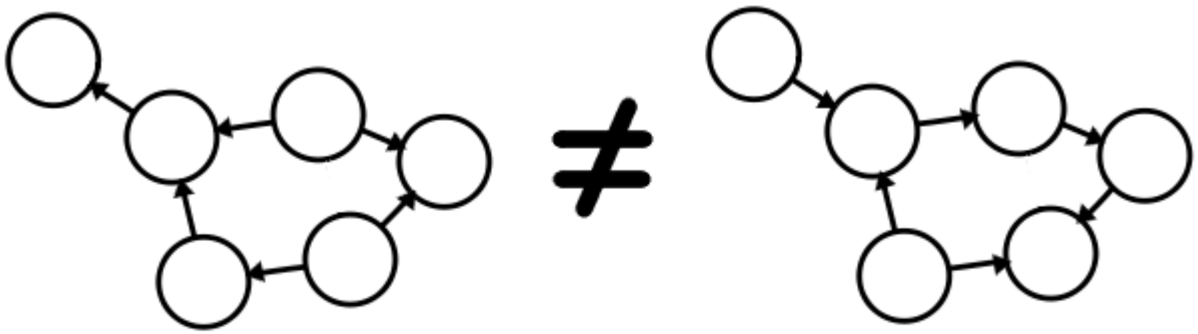
Currently, paths are the only measure

Not really expressive... *isn't there anything else*?

Directionality of edges

Asymmetric relations are frequent

But what do directions *mean*?



Hypothesis

The most basic asymmetry: Hierarchies

Knowledge does not get any simpler than that

Specialization → Generalization

Descendant → Ancestor

What do the descendants and ancestors of an entity tell us about that entity?

- *After meeting a thousand cats, what do you know about "**cat**"?*
- *After meet animals with claws, what do you know about "**cat**"?*
  - *Quite a lot actually...*

Overview

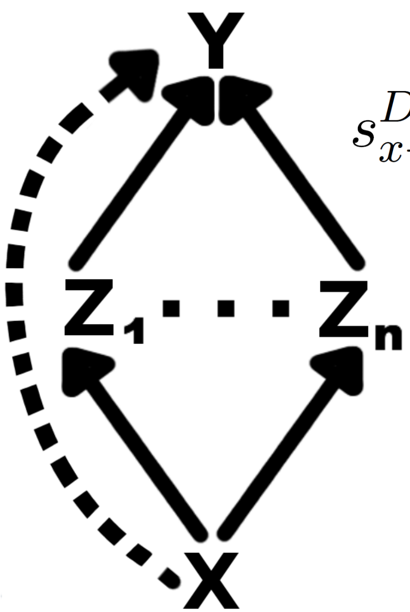
- ◎Motivation
- ◎State of the Art
- ◎Hypothesis
- ◎Hierarchical Link Prediction
- ◎Computational Models & Designs
- ◎Data Sets & Results
- ◎Conclusions
- ◎Discussion & Future Work

Hierarchical Link Prediction

◎The INFerence score: x→y?

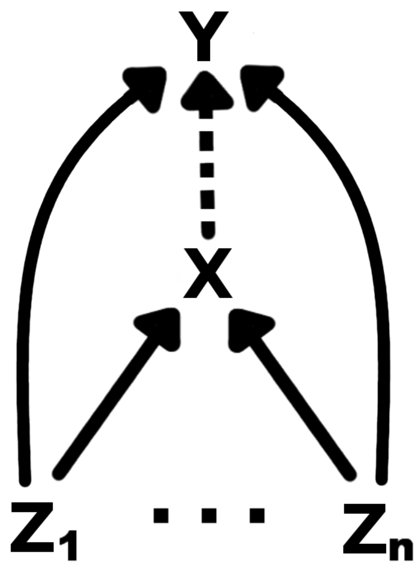
-Given the generalizations of x, A(x), is x→y coherent? *Deductive reasoning (DED)*

-Given the specializations of x, D(x), is x→y coherent? *Inductive reasoning (IND)*



$$s_{x \to y}^{DED} = \frac{|A(x) \cap D(y)|}{|A(x)|}$$

$$s_{x \to y}^{IND} = \frac{|D(x) \cap D(y)|}{|D(x)|}$$



Hierarchical Link Prediction

The INFerence score

Just add the evidence: INF = DED + IND

But INF is purely *proportional*:

$$s^{DED}_{x \to y} = \frac{|A(x) \cap D(y)|}{|A(x)|} \qquad s^{IND}_{x \to y} = \frac{|D(x) \cap D(y)|}{|D(x)|}$$

While all top scores are ***accumulative***:

$$s^{CN}_{x,y} = |\Gamma(x) \cap \Gamma(y)|$$

$$s^{AA}_{x,y} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(|\Gamma(z)|)}$$

$$s^{RA}_{x,y} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}$$

INFerence modifications

Accumulative scores: ***Skip low-degree vertices. Rich get richer***.
Proportional evidence is important too: ***Make it hybrid***

$$s_{x \to y}^{DED\_LOG} = \frac{|A(x) \cap D(y)|}{|A(x)|} * log(|A(x)|)$$

Deduction is more reliable: INF_2D = 2*DED + IND
INF_LOG, INF_LOG_2D a new family of hybrid scores

$$s_{x \to y}^{IND\_LOG} = \frac{|D(x) \cap D(y)|}{|D(x)|} * log(|D(x)|)$$

Overview

- ◎Motivation
- ◎State of the Art
- ◎Hypothesis
- ◎Hierarchical Link Prediction
- ◎**Computational Models & Designs**
- ◎Data Sets & Results
- ◎Conclusions
- ◎Discussion & Future Work

Computational Models & Designs

Similarity-based is scalable … enough?

Graph with 1M vertices → $1 \cdot 10^{12}$ similarities
Unfeasible to compute them one by one!

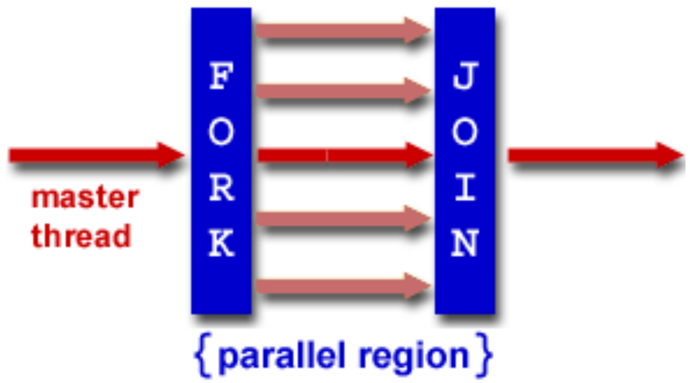
Similarity-based is parallelizable … how?

Very parallel... **embarrassingly** parallel!
Similarities are independent of one another
Parallel computing models are a must

Computational Models & Designs

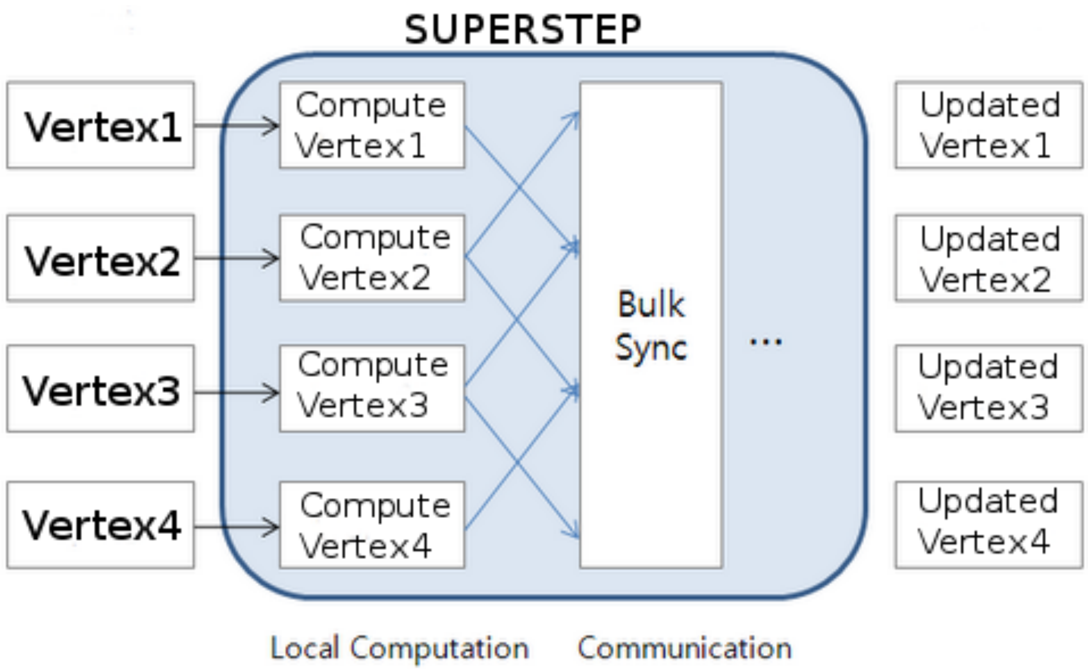
General parallel computing model

-Fork-join (OpenMP)
-Tested on MareNostrum (BSC)



Graph-specific parallel computing model

-Pregel (ScaleGraph)
-Tested on TSUBAME (UCD/JSTCrest)



Computational Models & Designs

Different algorithmic designs are possible

Intersection-based

$\forall\ v^1 \in \mathbf{N}$

- $\forall\ v^2 \in \mathbf{N}$
    - intersection(neigh(v^1),neigh(v^2))

Traverse-based

$\forall\ v \in \mathbf{N}$

- $\forall$ neigh(v)
    - $\forall$ neigh(neigh(v))

Computational Models & Designs

Intersection-based

-All v1,v2 paths found at the same time

-High complexity: $O(N^2 \cdot k)$

-High locality

Traverse-based

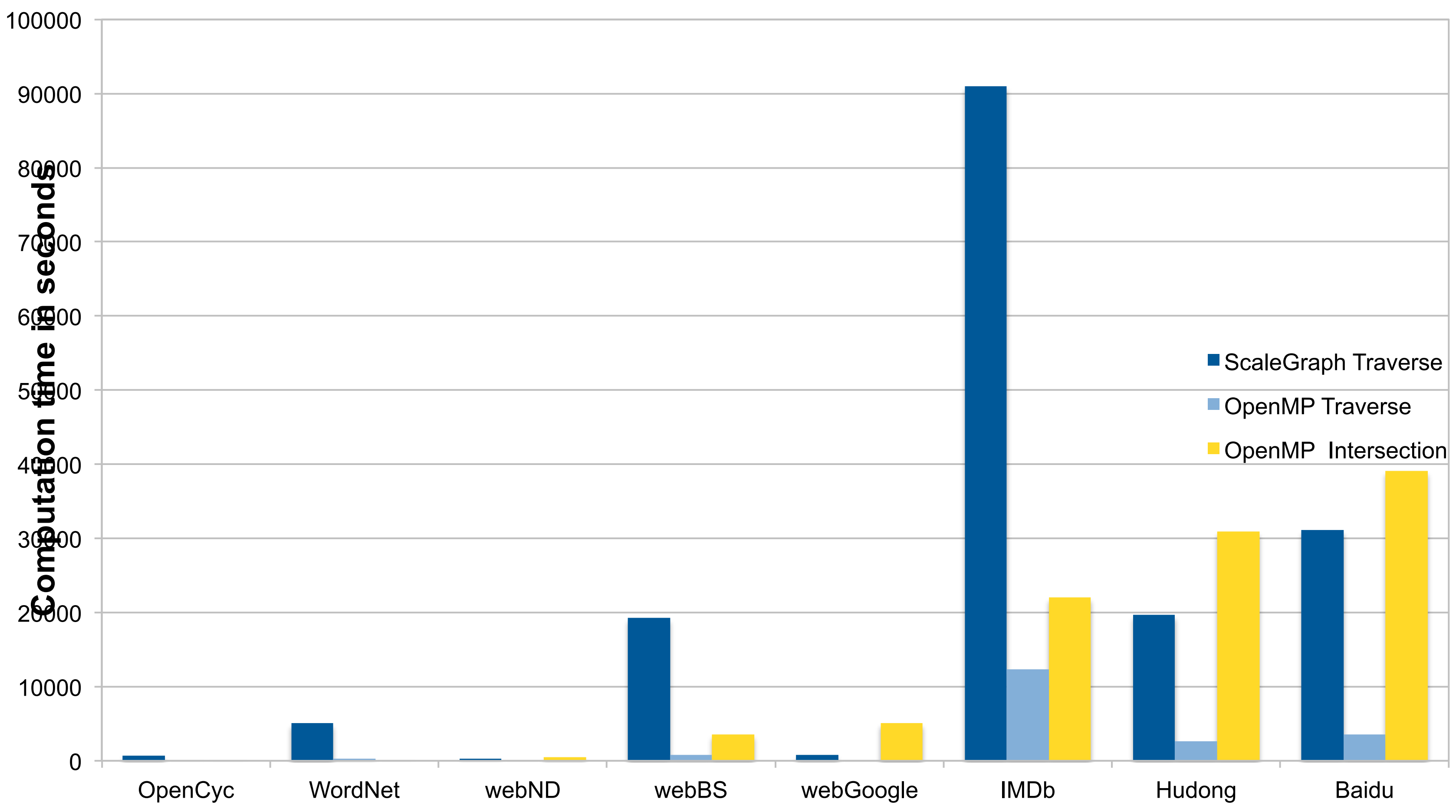
-v1,v2 paths found one at a time

-Low complexity: $O(N \cdot k^3)$

-No locality

Computational Models & Designs

Computation times of both designs

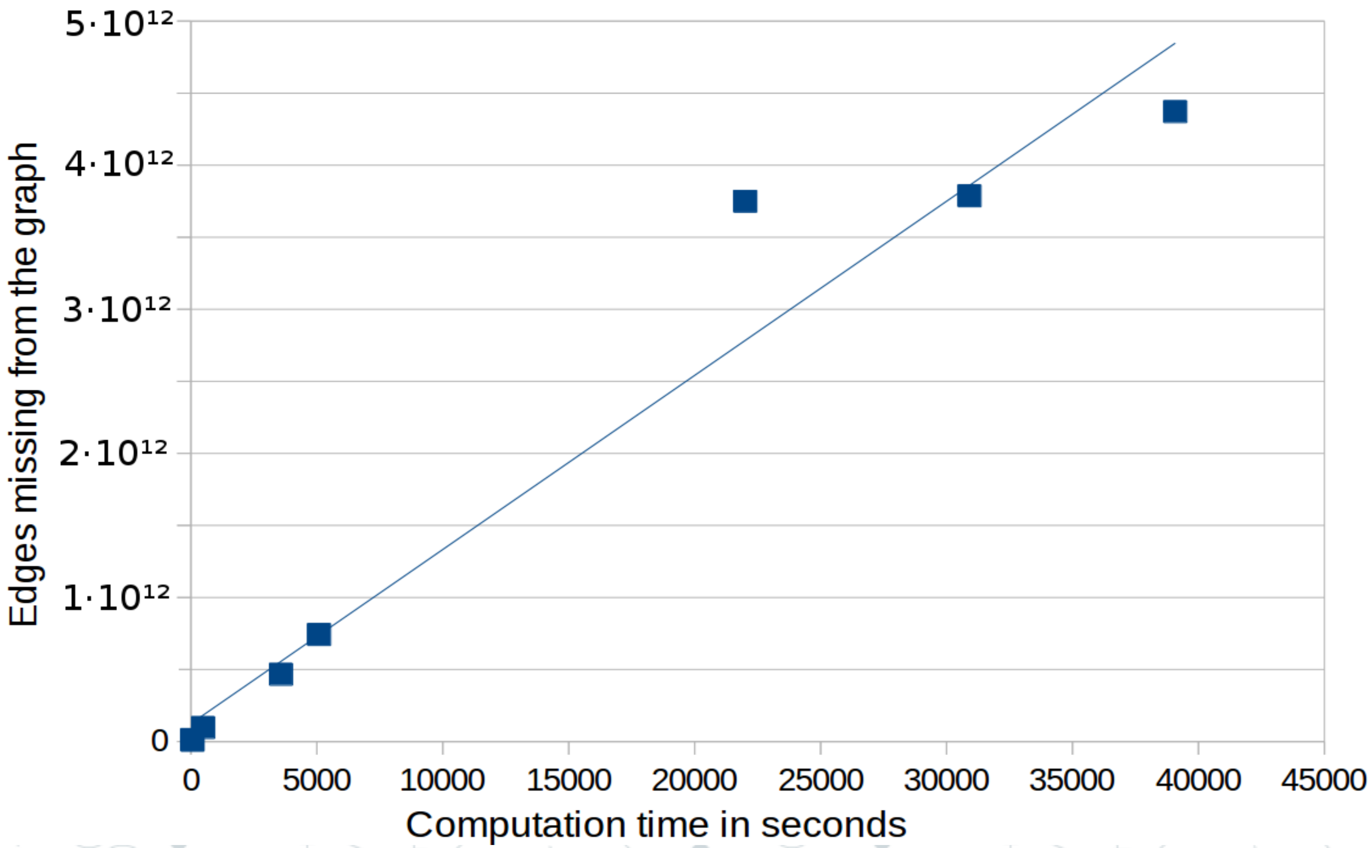


Computational Models & Designs

Intersection design: Good for superhubs (locality)

-Cost based on missing edges

OpenMP computation times and regression

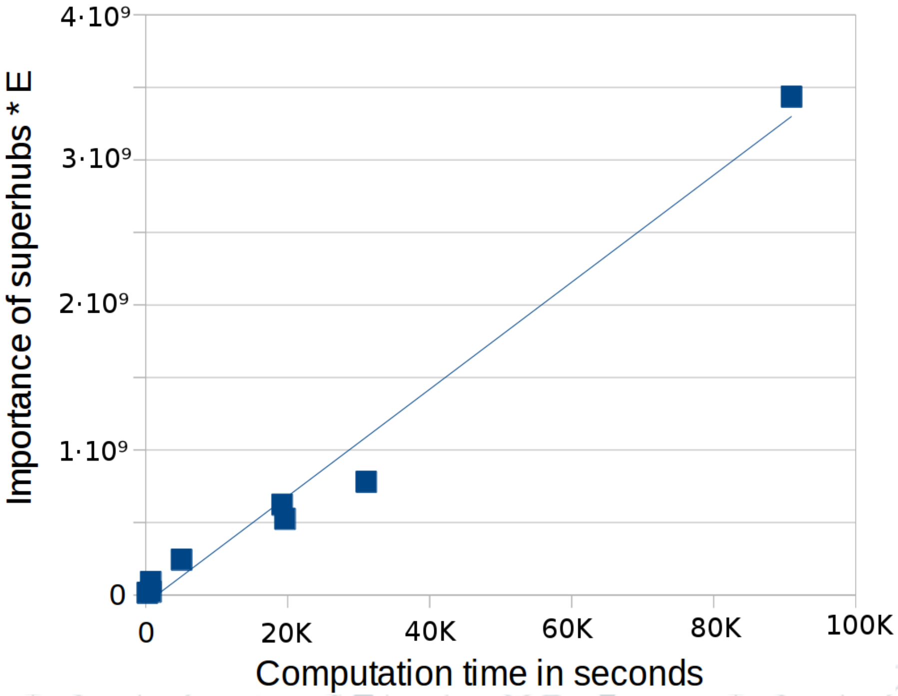


Computational Models & Designs

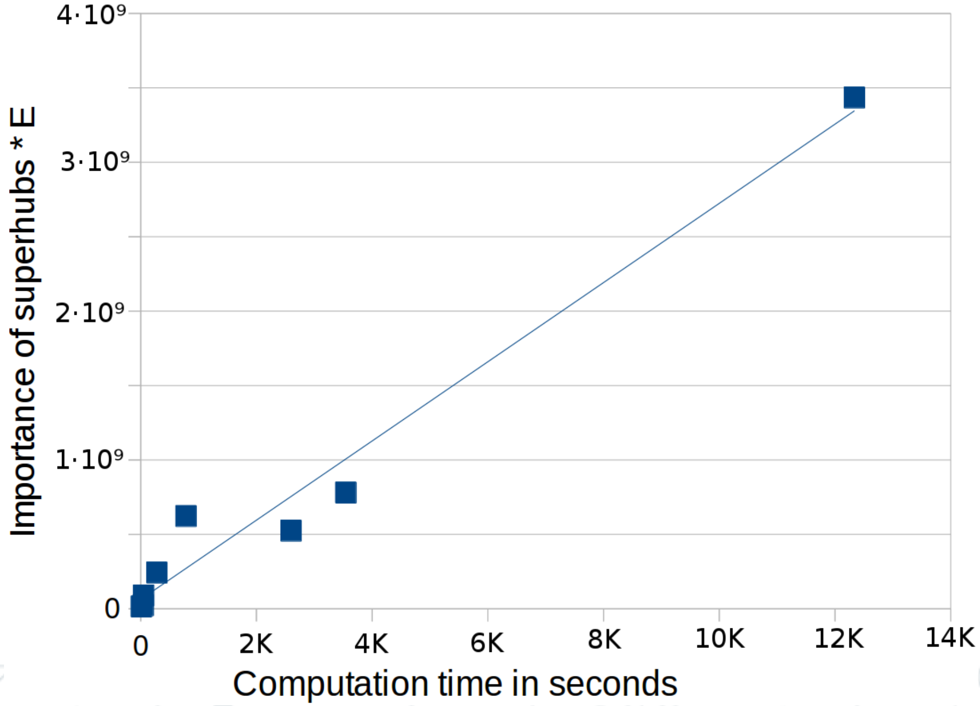
Traverse design: Good for all but superhubs (complexity)

-Cost based on graph size and superhubs relevance

ScaleGraph



OpenMP



Computational Models & Designs

OpenMP

-Control over data-structures (type, order)

ScaleGraph

-Designed for large-scale graphs

-Automatic management of data and communications

What is a small/large graph?

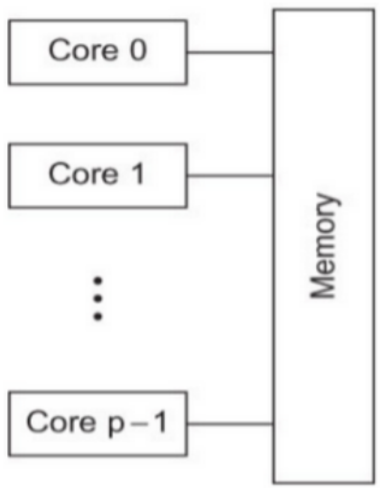
-Requires lots of memory

-Requires lots of computing units

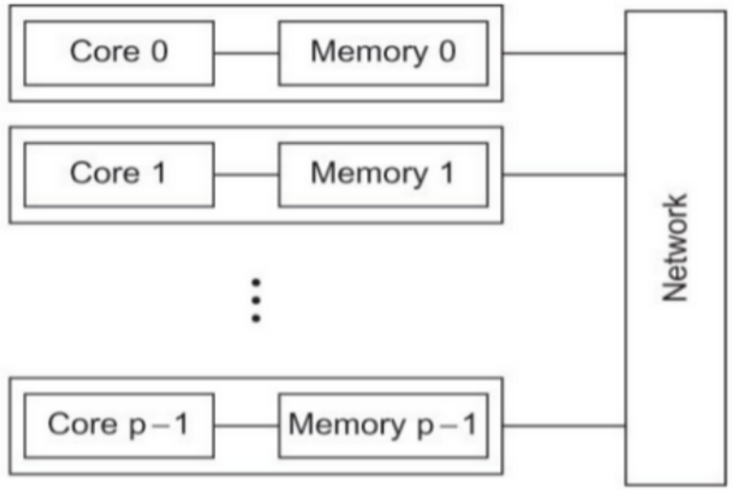
Computational Models & Designs

Single machines have a limit of memory and of computing units. Eventually...

Shared memory paradigm



Distributed memory paradigm



OpenMP/ScaleGraph | OmpSs/ScaleGraph

https://pm.bsc.es/ompss

Overview

- ◎Motivation
- ◎State of the Art
- ◎Hypothesis
- ◎Hierarchical Link Prediction
- ◎Computational Models & Designs
- ◎Data Sets & Results
- ◎Conclusions
- ◎Discussion & Future Work

Data Sets & Results

INF assumes hierarchical directionality... should work on hierarchical graphs

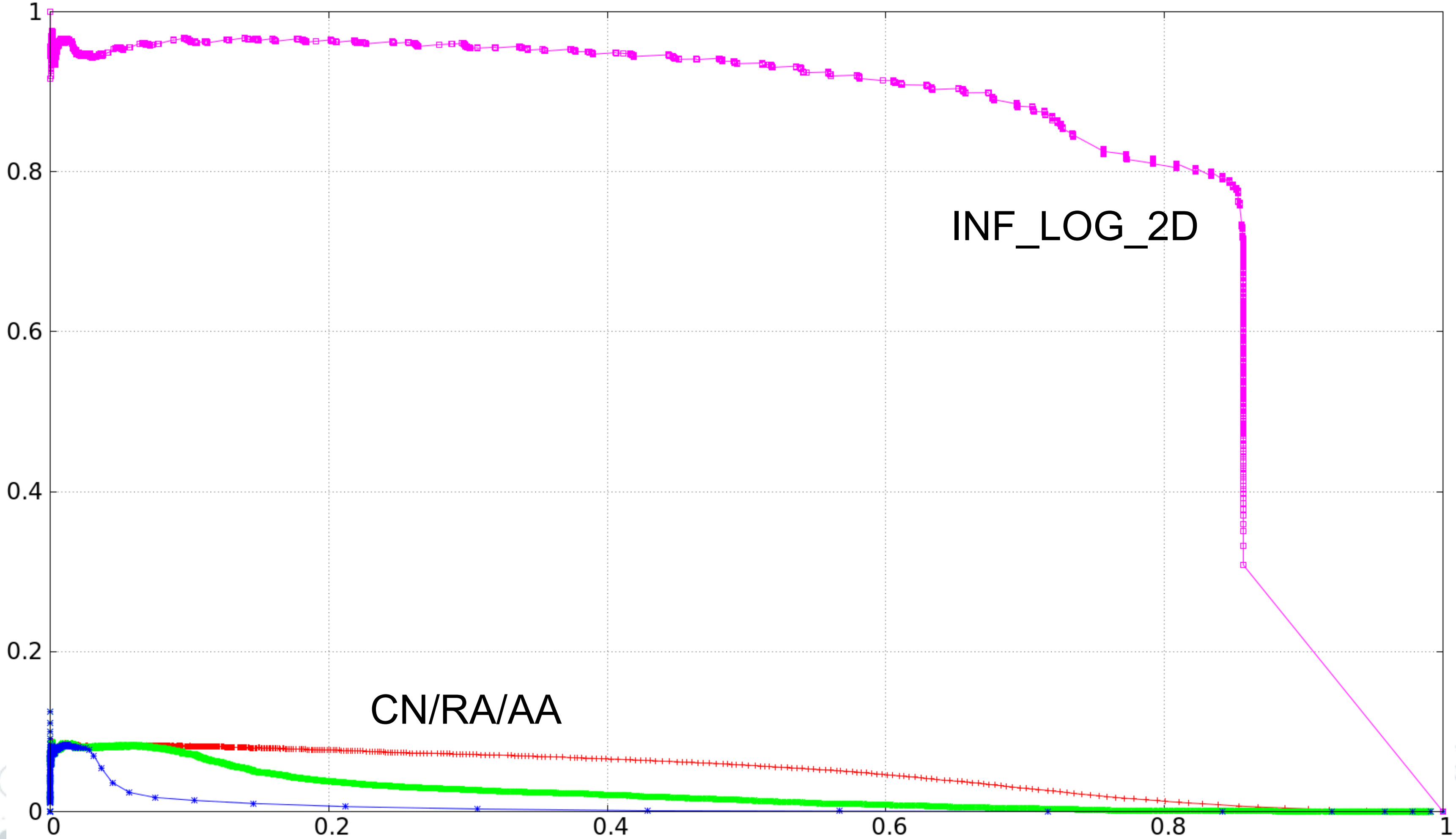
- Wordnet (lexical hyponym/hypernym)
  - 89K vertices, 698K edges
- OpenCyc (ontological subClass, instanceOf)
  - 116K vertices, 345K edges

Evaluation through AUC – Precision/Recall curves

- **Random remove of 10% for test**

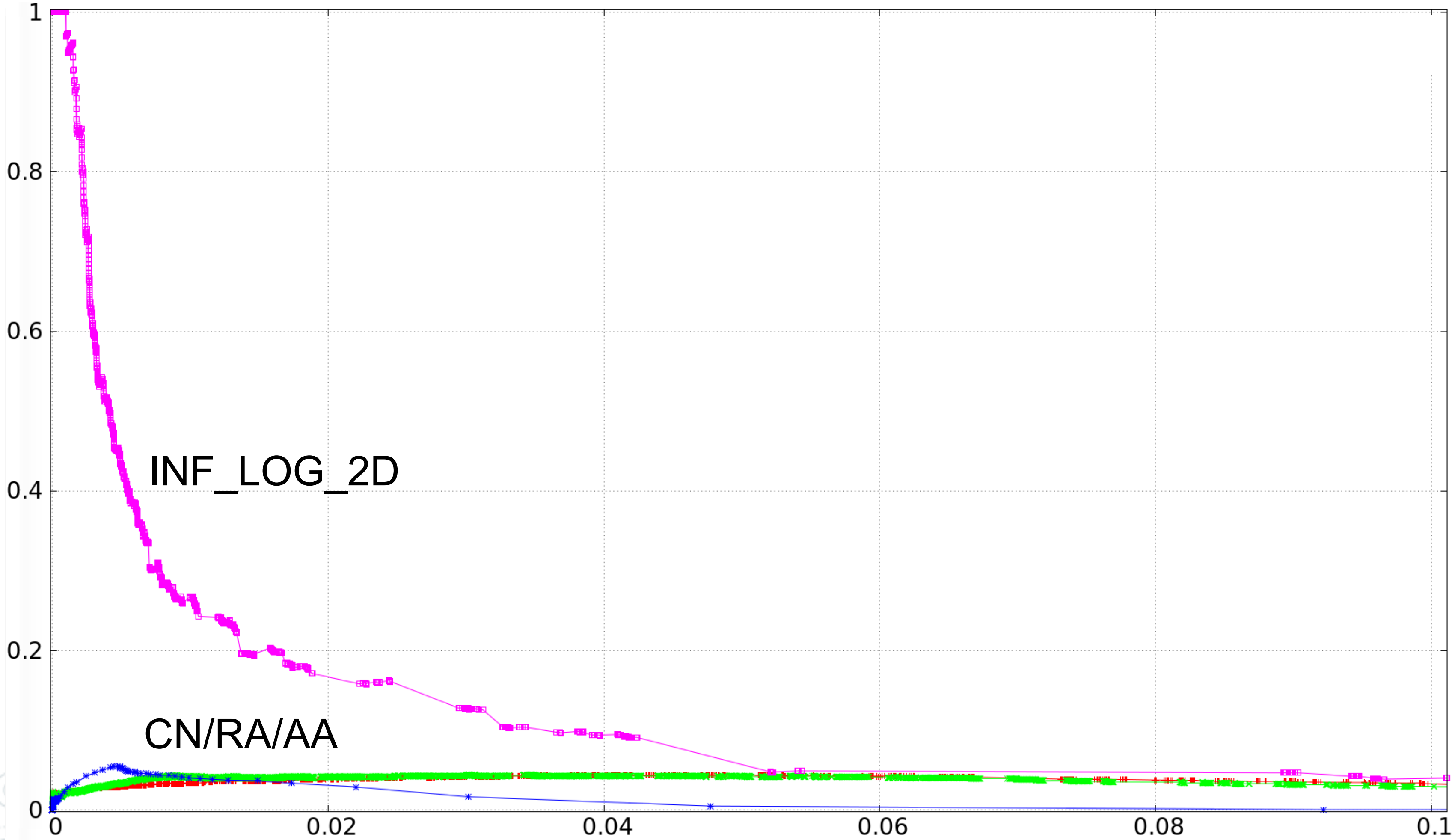
Data Sets & Results

◎WordNet – RA (red), AA (green) CN (blue), INF_LOG_2D (pink)



Data Sets & Results

OpenCyc – RA (red), AA (green) CN (blue), INF_LOG_2D (pink)

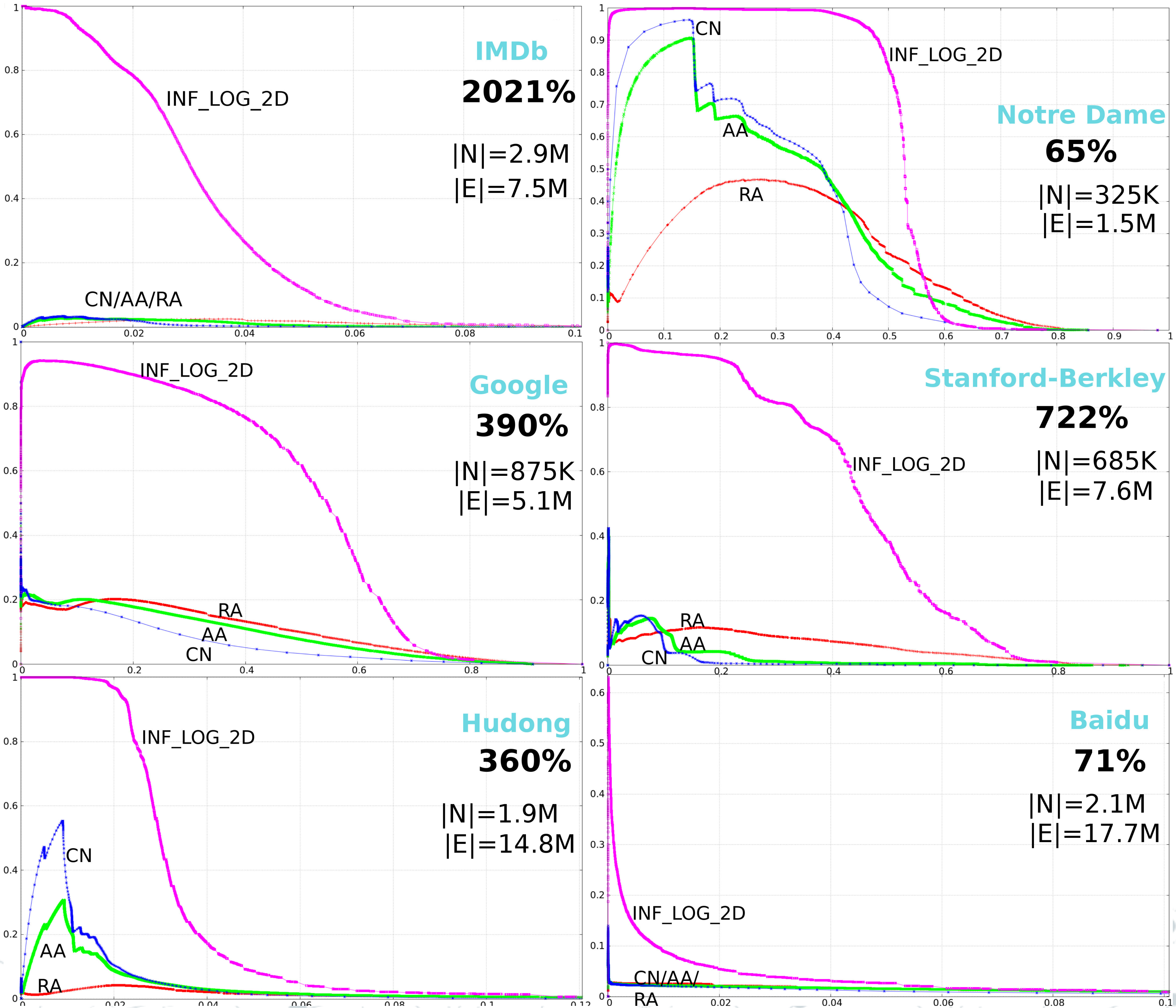


Data Sets & Results

So it work for hierarchical graphs... what about non-hierarchical ones?

- IMDb (movies, directors, genres and tags)
  - 1.9M vertices, 7.5M edges

- Web graphs* (web pages and hyperlinks)
  - Notre Dame: 325K vertices, 1.5M edges
  - Stanford-Berkley: 685K vertices, 7.6M edges
  - Google: 875K vertices, 5.1M edges
  - Hudong: 1.9M vertices, 14.8M edges
  - Baidu: 2.1M vertices, 17.7M edges


IMDb
2021%
|N|=2.9M
|E|=7.5M
INF_LOG_2D
CN/AA/RA
Notre Dame
65%
|N|=325K
|E|=1.5M
INF_LOG_2D
CN
AA
RA
Google
390%
|N|=875K
|E|=5.1M
INF_LOG_2D
RA
AA
CN
Stanford-Berkley
722%
|N|=685K
|E|=7.6M
INF_LOG_2D
RA
AA
CN
Hudong
360%
|N|=1.9M
|E|=14.8M
INF_LOG_2D
CN
AA
RA
Baidu
71%
|N|=2.1M
|E|=17.7M
INF_LOG_2D
CN/AA/
RA

Overview

◎Motivation
◎State of the Art
◎Hypothesis
◎Hierarchical Link Prediction
◎Computational Models & Designs
◎Data Sets & Results
◎Conclusions
◎Discussion & Future Work

Conclusions

- Hierarchies are latent in some large graphs

  -*“Naturally*!”

- Hierarchies can be used for Link Prediction

  -*“No they can't. They should!”*

- It is feasible to do large-scale Link Prediction

  -*“Link Prediction and HPC: a perfect couple”*

Conclusions

◎INFerence

-Do not build a model, just use it

-*Proportional-Accumulative* scores

-Huge leap in predictive performance

λ**Precision** **Scalability**

◎Evaluation under class super-imbalance

-Do not do it all, just do it right

Overview

Discussion & Future Work

- Data-intensive tasks: Cost, data structures and locality
- Large-scale graphs
  - OmpSs/Scalegraph on cluster
- HPC & Graph Mining: Models, algorithms, …
- Traverse vs intersection design

Discussion & Future Work

- Applications
  - Search engines, product recommendation, research support, etc.
- Improving INFerence
  - Tunned parameters
  - Quasi-local INF
- Deep Learning + Graph Mining

Thanks

KEMLg
BSC

dariog@lsi.upc.edu

dario.garcia@bsc.es

ia@cs.upc.edu

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Simple line icons by Mirko Monti
- E-commerce icons by Virgil Pana
- Streamline iconset by Webalys
- Presentation template by SlidesCarnival
- Photographs by Unsplash & Death to the Stock Photo (license)