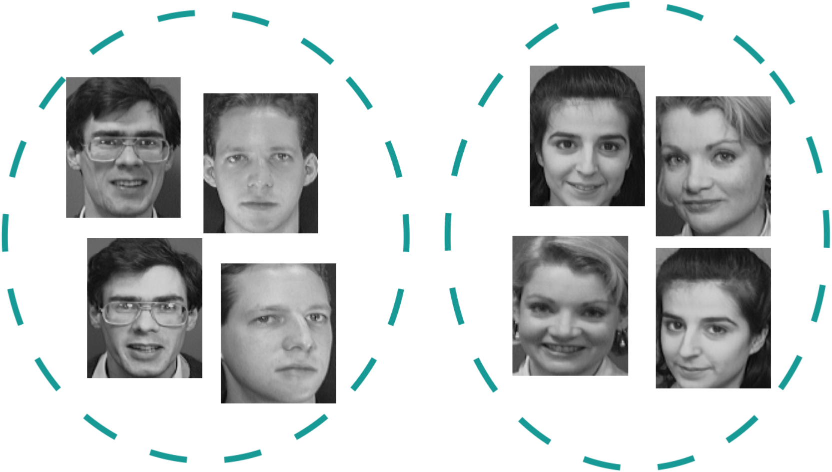
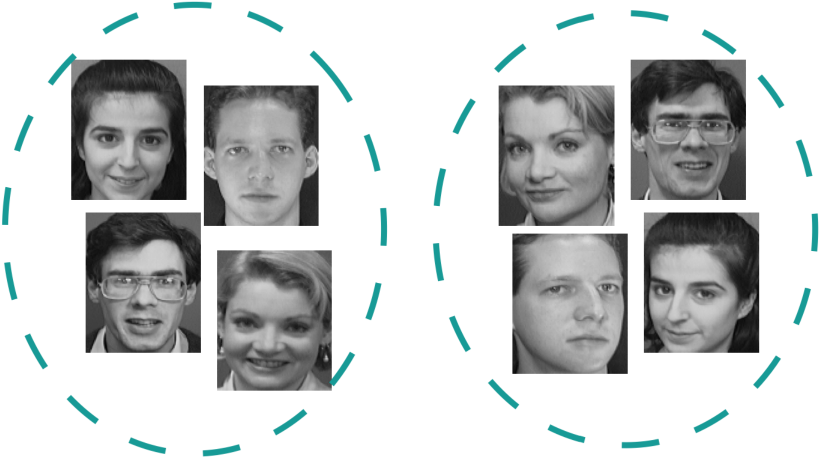
An expressive dissimilarity measure for relational clustering using neighbourhood trees

Sebastijan Dumančić, Hendrik Blockeel

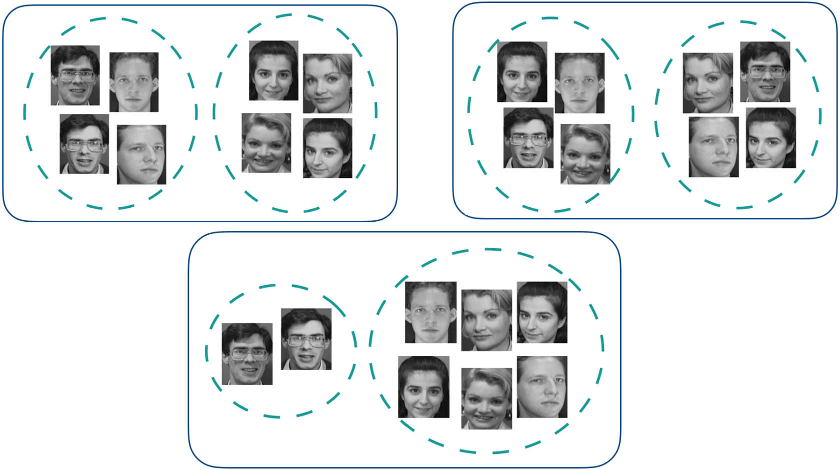
DTAI, CS Department, KU Leuven

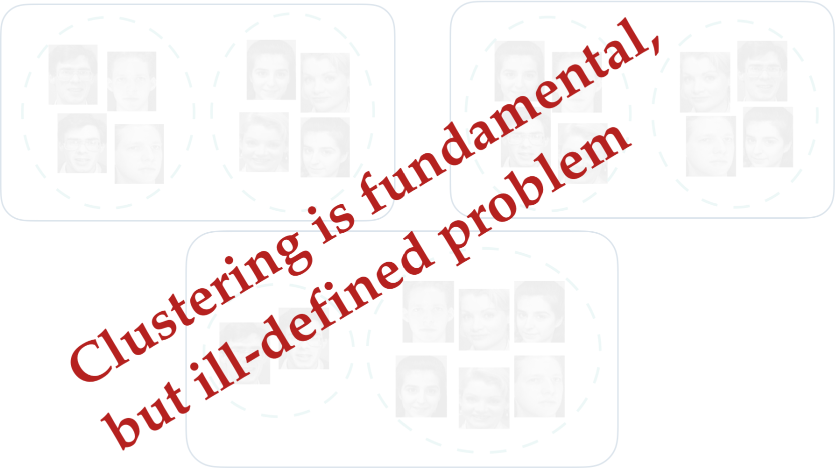
ECML PKDD 2017, Journal track





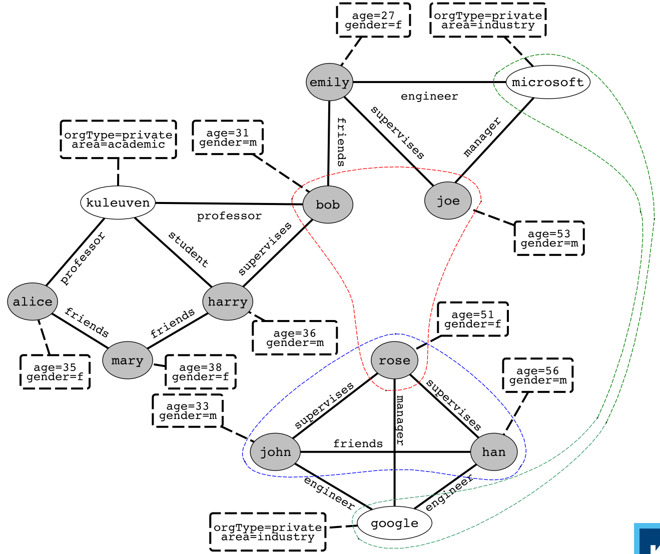
1 – Identifying groups in data





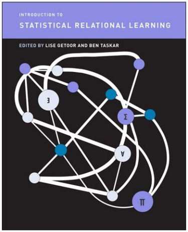
Clustering is fundamental, but ill-defined problem

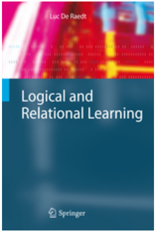
1 – What about relational data?



Machine learning with a powerful knowledge representation language

- usually based on first-order logic





Common representation for:

- vectors
- graphs
- sequences
- ...

... with a unifying reasoning and learning engine a)

```
person(bob,25,m,msc)
person(emily,27,f,msc)
organization(kuleuven,private,academic)
organization(microsoft,private,industry)
friends(bob,emily)
friends(emily,bob)
works_for(bob,kuleuven,professor,2000)
works_for(emily,microsoft,engineer,2300)
```

b)

person

PName	Age	Gender	Education
bob	25	m	msc
emily	27	f	msc

organization

OName	OrgType	Area
kuleuven	private	academic
microsoft	private	industry

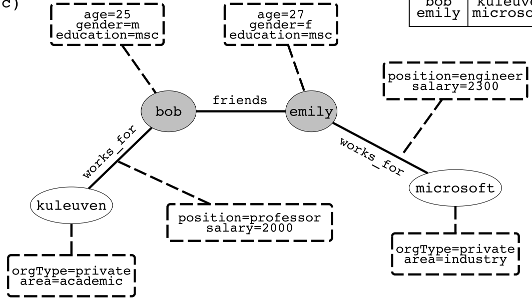
works_for

PName	OName	Position	Salary
bob	kuleuven	professor	2000
emily	microsoft	engineer	2300

friends

PName1	PName2
bob	emily
emily	bob

c)



a)

person(bob,25,m,msc)
person(emily,27,f,msc)
organization(kuleuven,private,academic)
organization(microsoft,private,industry)
friends(bob,emily)
friends(emily,bob)
works_for(bob,kuleuven,professor,2000)
works_for(emily,microsoft,engineer,2300)

b)

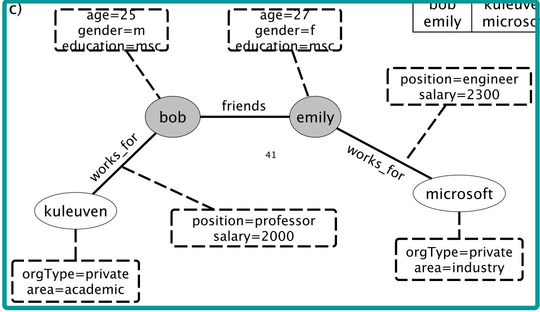
person			
PName	Age	Gender	Education
bob	25	m	msc
emily	27	f	msc

organization		
OName	OrgType	Area
kuleuven	private	academic
microsoft	private	industry

works_for			
PName	OName	Position	Salary
bob	kuleuven	professor	2000
emily	microsoft	engineer	2300

friends	
PName1	PName2
bob	emily
emily	bob

c)



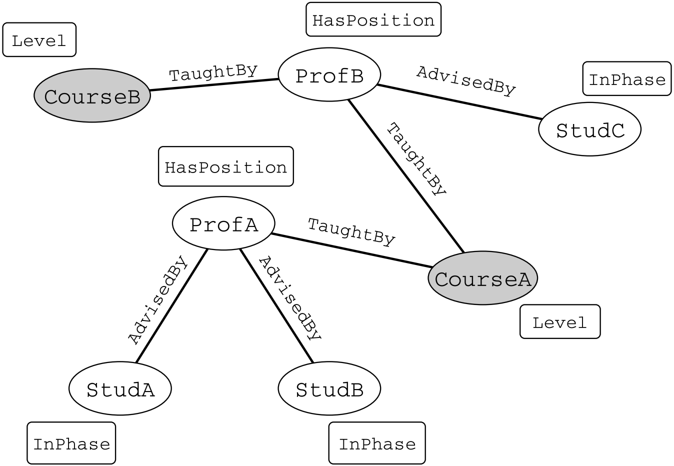
Hybrid similarities	Graph kernels	Relational similarities
incorporate link information into attribute-based similarity	structural similarities of graphs	comparing logical constructs
measure the similarity of connected vertices	*random walks, propagation of information*	*logical formulas in common, matching terms*

Hybrid similarities	Graph kernels	Relational similarities
incorporate link information into attribute-based similarity	structural similarities of graphs	comparing logical constructs
measure the similarity of connected vertices	*random walks, propagation of information*	*logical formulas in common, matching terms*

Impose a fixed bias

1. Overture
2. How do we do it now?
3. An expressive dissimilarity for relational data
4. Experiments and results
5. Summary

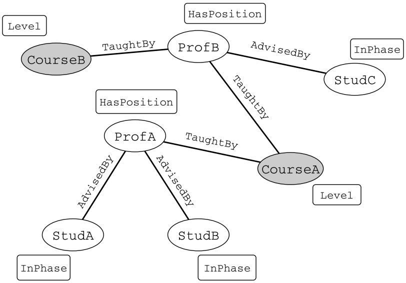
3 – How similar are ProfA and ProfB?



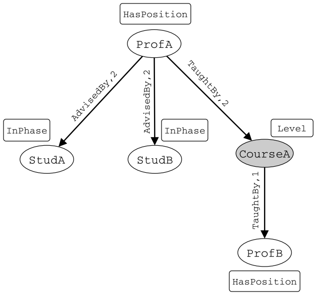
A similarity measure for relational data should:

- incorporate multiple views of similarity
- be easily adaptable
- take attributes and relationships into account
- insensitive to neighbourhood size
- be efficient

Neighbourhood trees summarize the neighbourhood of an instance/example

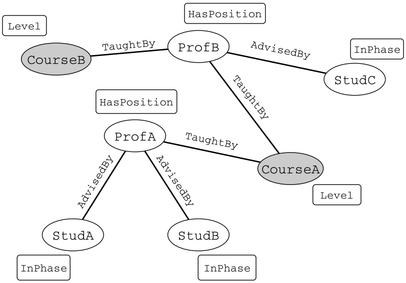


Data

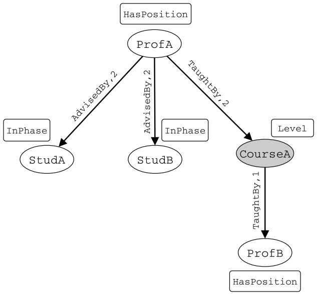


Neighbourhood tree

Neighbourhood trees summarize the neighbourhood of an instance/example



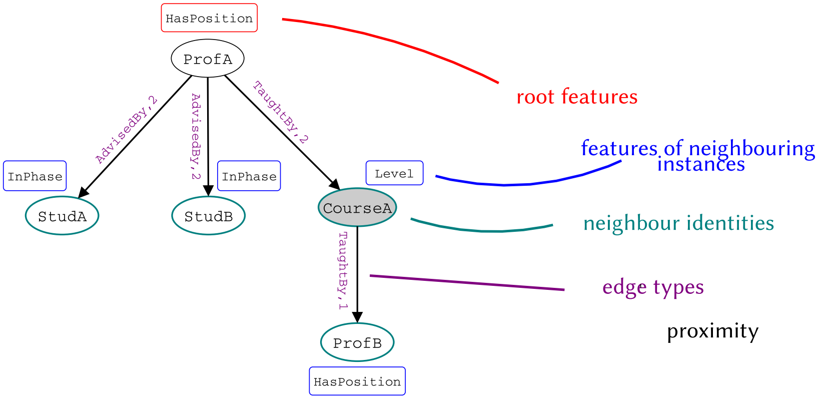
Data



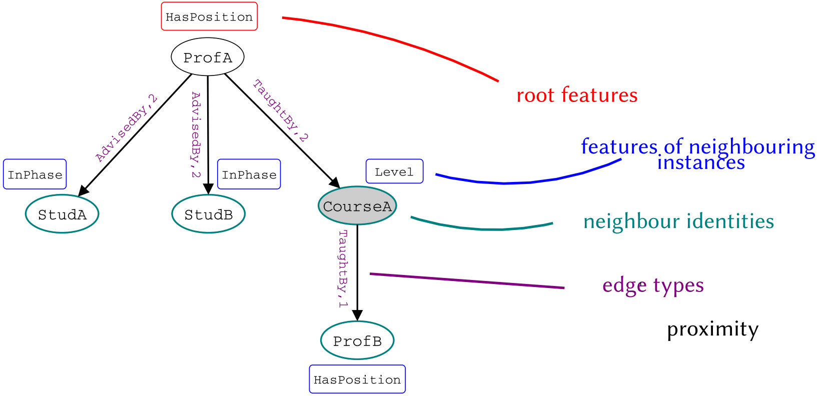
Neighbourhood tree

Similarity of instances = similarity of their neighbourhood trees

Decompose NTs into semantic parts

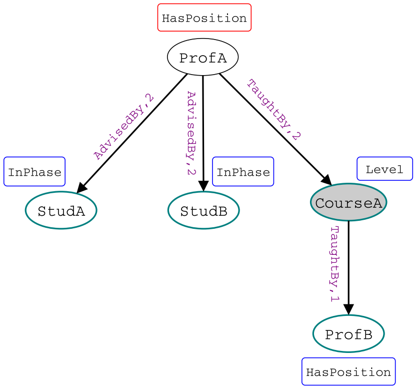


Decompose NTs into semantic parts



similarity = linear combination of *similarities of individual semantic parts*

$$(w_1, w_2, w_3, s_4, w_5)$$



Decompose NT in multisets of:

- attribute
- edge labels
- vertex identities

per level and vertex type

Multiset of edge labels (level 1):
{ (Advised,2), (Advised,2), (TaughtBy,2) }

Compare two multisets, A and B with χ^2 distance

$$\chi^2(A,B) = \sum_{x \in A \cup B} \frac{(f_A(x) - f_B(x))^2}{f_A(x) + f_B(x)}$$

Many of the existing similarities are a special case:

- hybrid similarities
- relational similarities

... or they can be defined over neighbourhood trees (graph kernels) with different biases:

- makes it easier to compare the imposed biases

Many of the existing similarities are a special case:

- hybrid similarities
- relational similarities

... or they can be defined over neighbourhood trees (graph kernels) with different biases:

- makes it easier to compare the imposed biases

Additionally: effective - linear in the number of unique elements in a multiset

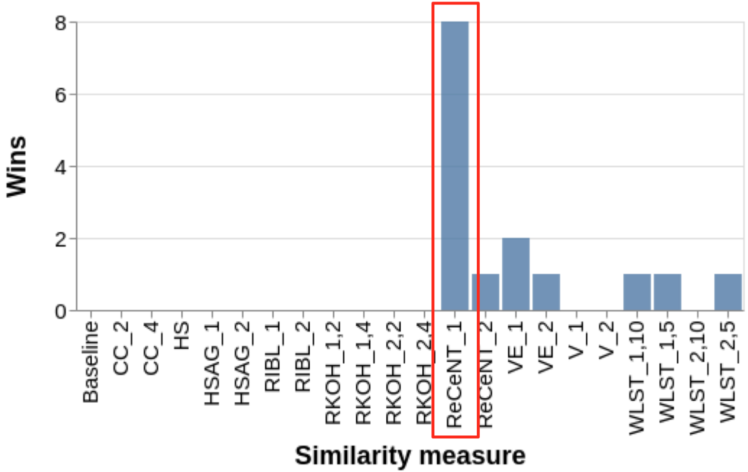
Datasets:

- IMDB
- UWCSE
- Mutagenesis
- WebKB
- TerroristAttacks

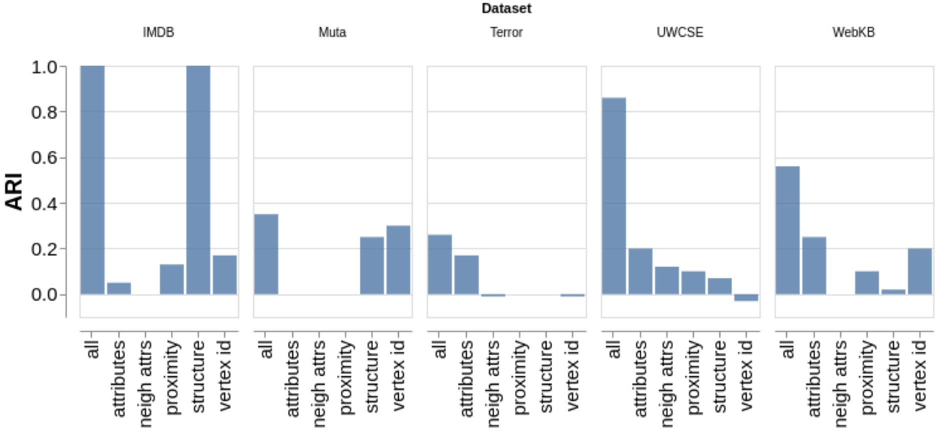
Questions:

- *Quality of the obtained clustering?*
- *Are different views really necessary?*
- *Can we learn the bias from data?*
- *Can we learn the bias from labels?*

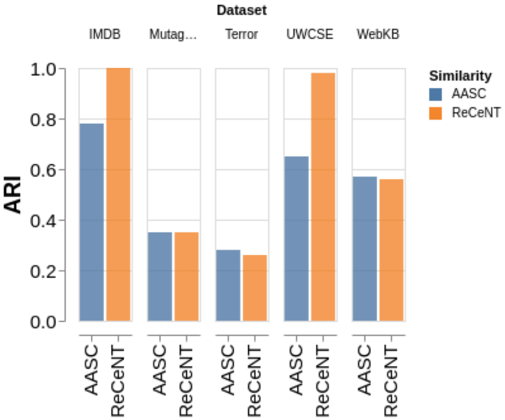
- combined with spectral and hierarchical clustering
- a wide range of existing similarity measures
- performance measure: ARI/Accuracy



Takeaway message: incorporating multiple biases consistently performs well



Takeaway message: relational data requires multiple views of similarity in order to find informative clusters



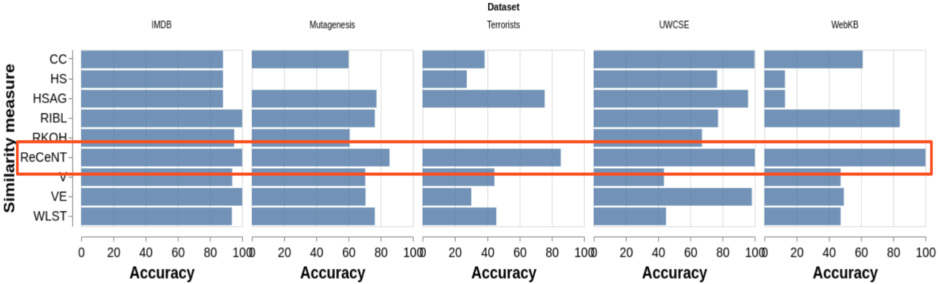
`ReCeNT` with $w_i = 0.2$
vs.
AASC + `ReCeNT`

AASC - given multiple similarity matrices, find an optimal combination for clustering

barely any benefit

Huang, Chuang, Chen: Affinity Aggregation for Spectral Clustering

Similarity measure in combination with a kNN (parameters optimised with CV)



Takeaway message: when labels are provided, `ReCeNT` outperforms the competing similarities

A similarity measure for relational data that:

- is versatile (meta-similarity)
- easily adaptable
- efficient
- generalization of many existing structured/relational sims
- works well across many different tasks

A similarity measure for relational data that:

- is versatile (meta-similarity)
- easily adaptable
- efficient
- generalization of many existing structured/relational sims
- works well across many different tasks

Code: `https://dtai.cs.kuleuven.be/software/recent`

S. Dumancic, H. Blockeel: *Clustering-Based Unsupervised Relational Representation Learning with an Explicit Distributed Representation*, IJCAI '17

S. Dumancic, H. Blockeel: *Demystifying Relational Latent Representations*, ILP '17