Refining Large Identity Graphs using the Unique Name Assumption

SHUAI WANG

JOE RAAD

PETER BLOEM

FRANK VAN HARMELEN

université PARIS-SACLAY

VU VRIJE UNIVERSITEIT AMSTERDAM

ESWC'23, Hersonissos, Greece

Identity, non-identity & near-identity

- Co-referring relationship: when two entities are the same (owl:sameAs)

*Bill said **Alice** would arrive soon, and **she** did.*

- Non-identity: when two entities are distinctly different

*Bill said **Alice** would arrive soon, then **Jane** arrived.*

- Near-identity: when two entities share most but not all feature values

***The United States** has officially restored diplomatic relations with Yugoslavia . . .*
***The White House** said the United States will provide 45 million dollars in food aid.*

Marta Recasens, Eduard Hovy, and M. Antònia Martí. "Identity, non-identity, and near-identity: Addressing the complexity of coreference." Lingua 121, no. 6 (2011): 1138-1152.

20

Marieke van Erp

ESWC'23 keynote

université PARIS-SACLAY

VU

IDENTITY CRISIS IN THE SEMANTIC WEB

- The sameas.cc identity graph:

  > extracted from 2015 LOD Laundromat crawl (38B triples in 650K RDF files)

  > 558M triples & 179M entities

  Connected components (CC)

  > 49M Connected Components (CCs)

  > the largest CC: 178K entities

  Error rate

  > error rate estimated around 3%[1], around 20% [2]

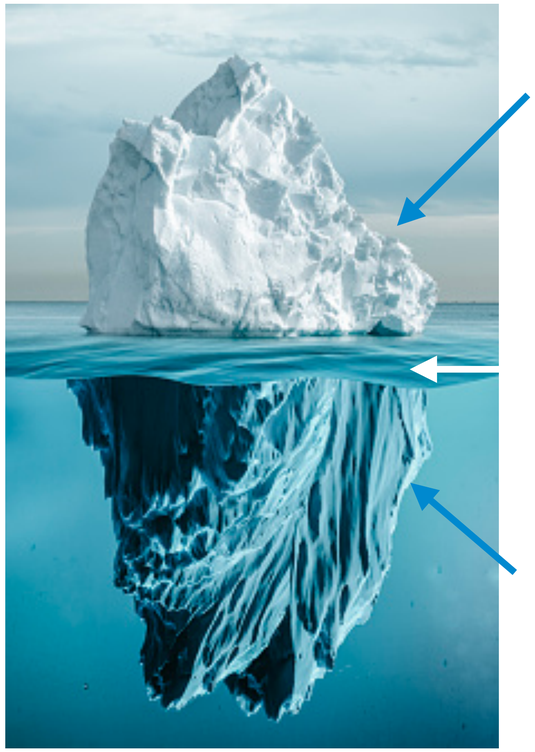
  > MetaLink (error degree) using the Louvain algorithm

- UNA: the Unique Name Assumption

[1] Aidan Hogan et al. “Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora”.

[2] Joe Raad. “Identity Management in Knowledge Graphs”. doctoral dissertation. PhD thesis. University of Paris-Saclay, 2018.

RELATED WORK & MOTIVATION &



Related work:

1) Content-based approach would fail at this scale

2) Graph/network algorithm: new attempts that could be evaluated

3) Inconsistency-based approach worth more exploration

Motivation:

Discovering new information that can be used for refinement

Semantic web is more than just nodes and links but semantics!

université PARIS-SACLAY

VU

RESEARCH QUESTIONS

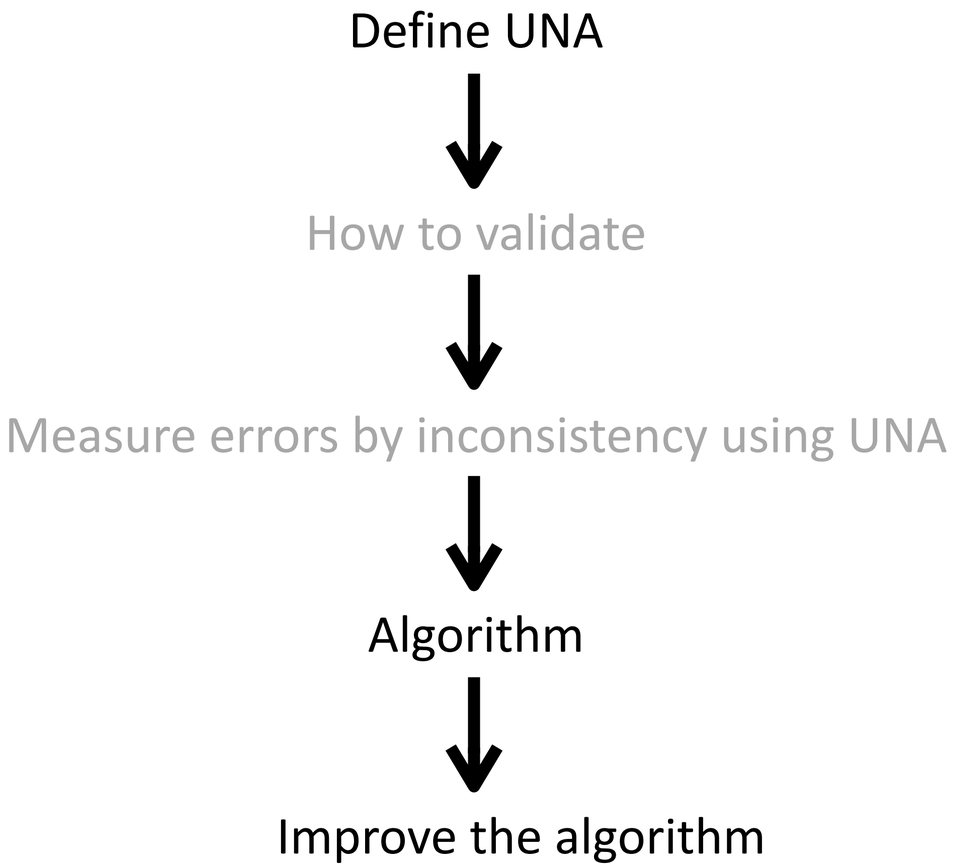
RQ1: How can we define a UNA for large integrated knowledge graphs?

RQ2: How do we validate various definitions of the UNA?

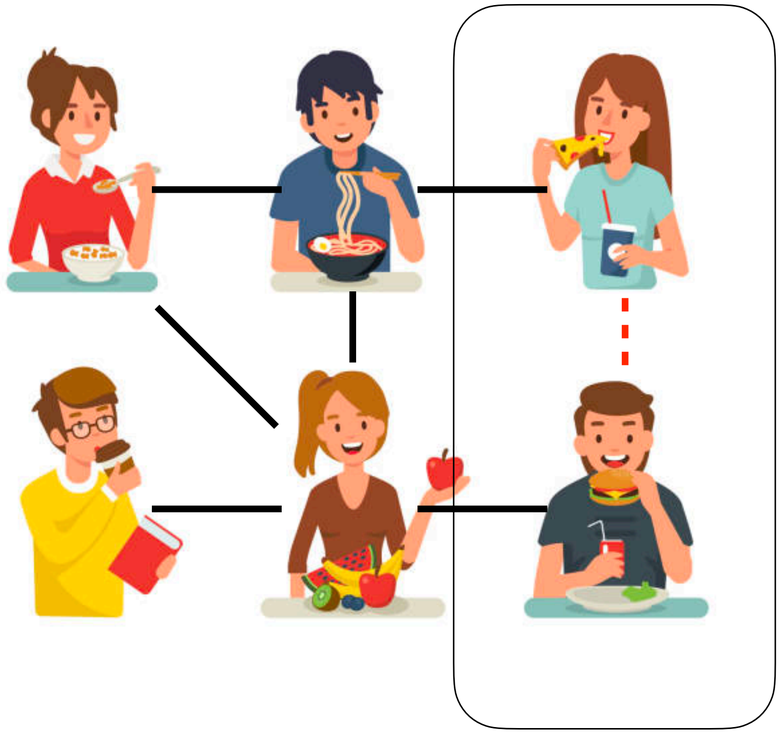
RQ3: Can the UNA give a reliable indication of errors in practice?

RQ4: Can we develop an efficient UNA-based algorithm for refinement?

RQ5: Is it possible to improve the results using additional information?



INCONSISTENCY & UNIQUE NAME ASSUMPTION



The Unique Name Assumption (UNA) supposes that two terms with distinct identifiers from the same knowledge base do not refer to the same real-world entity.

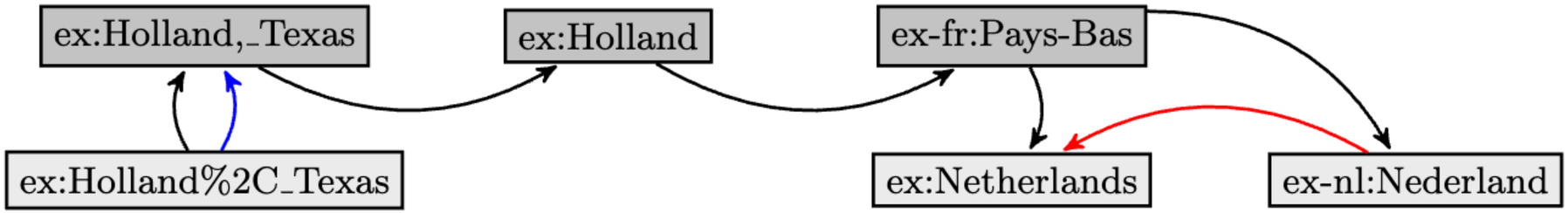
université PARIS-SACLAY

VU

UNIQUE NAME ASSUMPTION

UNA fails for large knowledge bases with redundant IRIs that capture

- various encodings
- languages
- namespaces
- versions
- letter cases



Semantic web has some network dynamics

A NEW UNA DEFINITION

Internal Unique Name Assumption (iUNA)

UNA + the following exceptions:

- redirection (version, namespace updates, dynamics, letter cases)
- a dead node, not found, unresolvable, redirects until reaching some error or has a timeout error while resolving
- percent encoded/decoded

V.S.

Naive Unique Name Assumption (nUNA) [3]

Quasi Unique Name Assumption (qUNA) [4]

université PARIS-SACLAY

[3] Andre Valdestilhas et al. “CEDAL: Time-Efficient Detection of Erroneous Links in Large-Scale Link Repositories”.
[4] Gerard de Melo. “Not Quite the Same: Identity Constraints for the Web of Linked Data

VU

ALGORITHM

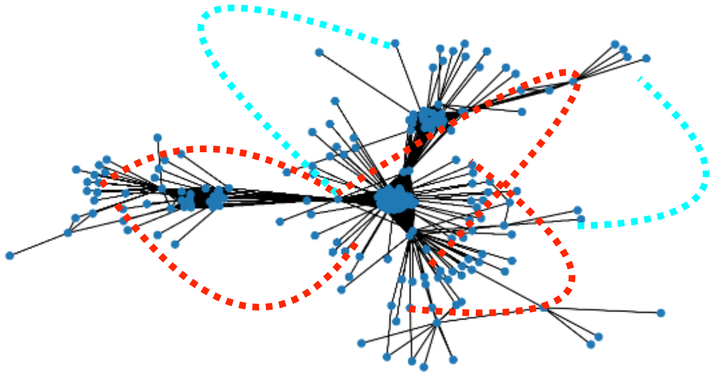
Provenance:

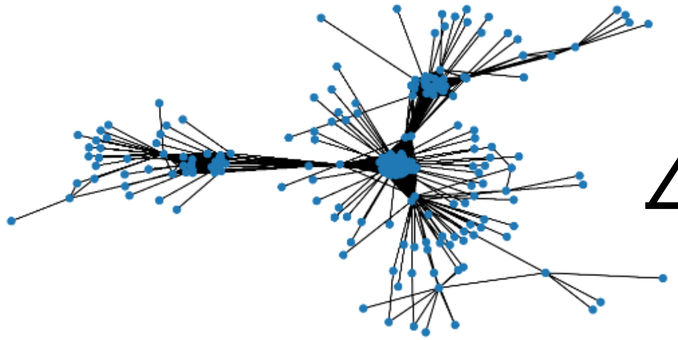
- rdfs:isDefinedBy
- rdfs:label
- rdfs:comment

Sample some pairs

Collect pairs violating UNA

Find shortest path between such pairs





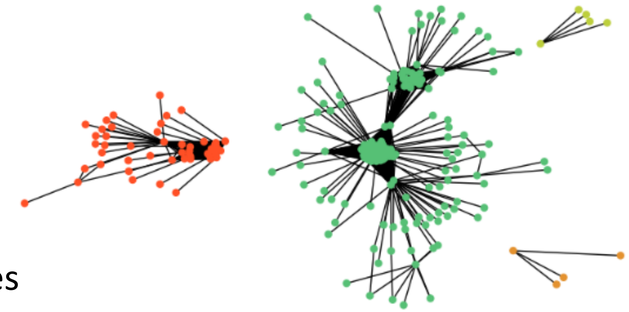
Input graph (a CC)

Encode each edge with propositional variable

Encode paths as "at least one edge is wrong"

SMT solver

Decode edges to remove



Output graph (a CC)

université PARIS-SACLAY

VU

GOLD STANDARD

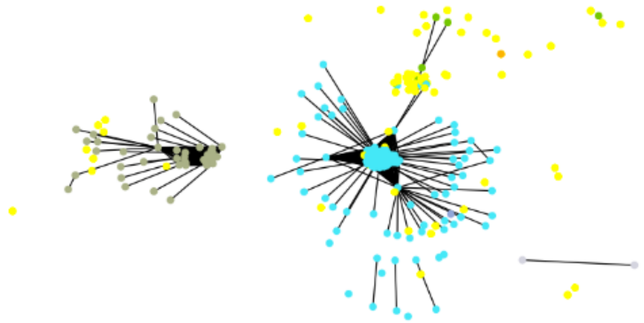
Gold standard

manually annotated IRIs from 28 CCs (max 1K nodes each)

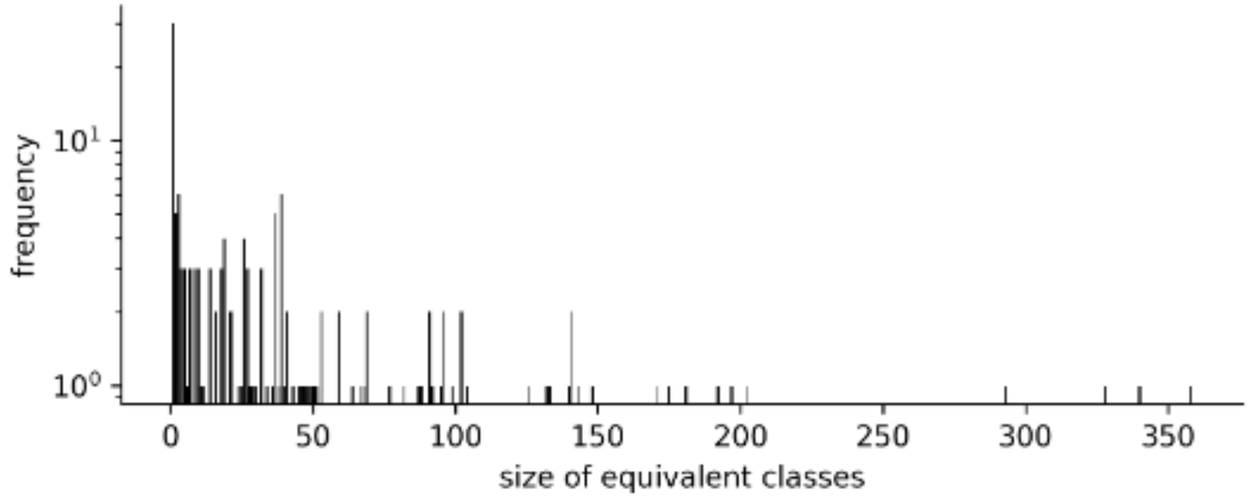
8,394 entities (with 232,311 links)

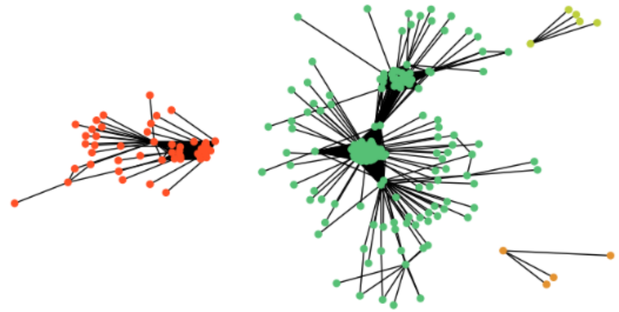
987 entities (11.75%) annotated as `unknown'.

estimate the error rate to be between 1.58% and 9.98%.



Manually annotated
gold standard (yellow = unknown)



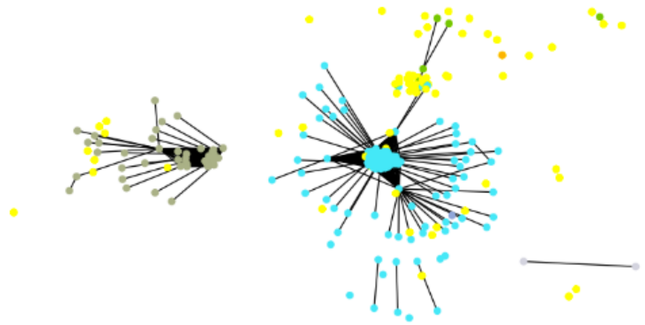


université PARIS-SACLAY

VU

EVALUATION

Precision and recall is not enough.



$$\Omega(G') = \sum_{C \in G'_{ccs}} \sum_{Q_e \in E(C)} \frac{|Q_e|}{|V|} \frac{|Q_e|}{|O_e|} \frac{|Q_e|}{|C|}.$$

		Evaluation set			
		precision	recall	Ω	$\lvert A \rvert$
Louvain	res=0.01	0.042	**0.727**	0.087	42,424.2
	res=1.0	0.042	0.660	0.084	43,610.0
Leiden		0.068	0.323	0.439	2,782.6
MetaLink	t=0.9	0.086	0.032	0.524	337
	t=0.99	0.013	0.001	0.635	99
nUNA	label, w1	0.042	0.063	0.597	684.6
	label, w2	0.061	0.075	0.580	697.4
	comment, w1	0.098	0.040	0.618	356.4
	comment, w2	0.063	0.036	0.606	431.2
qUNA	w_1	0.058	0.036	0.662	706.4
	w_2	0.101	0.054	**0.671**	634.2
iUNA	label, w1	0.122	0.013	0.652	236.8
	label, w2	**0.136**	0.028	0.647	235.0
	comment, w1	0.097	0.002	0.636	141.2
	comment, w2	0.117	0.003	0.638	173.8
Best performance		0.122		0.671	

université PARIS-SACLAY

VU

IMPROVING THE RESULTS WITH ADDITIONAL INFO

S: Significance: How many times it appears in the files of LOD Laundromat.

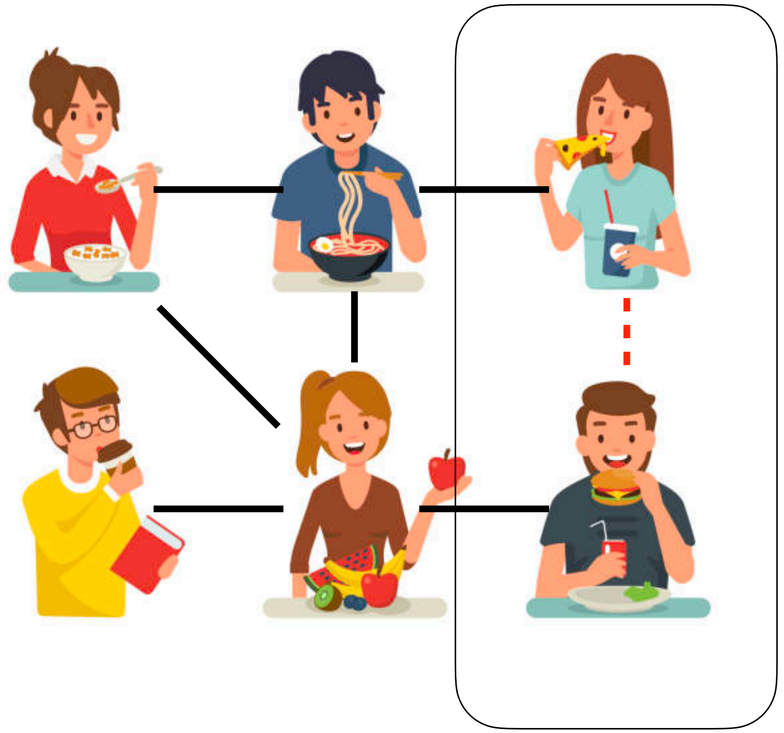
D: Among the erroneous edges in the gold, 38% involve at least one entity about disambiguation

After removing 501 disambiguation entities, the largest connected component is reduced from 178K to 83K entities (a reduction of 53.4%).

		Evaluation set			
		precision	recall	Ω	$\lvert A \rvert$
qUNA	w_1	0.058	0.036	0.662	706.4
	w_1^S	0.052	0.046	0.658	683.4
	w_1^D	0.044	0.051	0.683	738.4
	w_1^{SD}	0.039	0.068	0.662	682.2
	w_2	0.101	0.054	0.671	634.2
	w_2^S	0.042	0.034	0.668	645.6
	w_2^D	0.107	**0.077**	**0.675**	658.8
	w_2^{SD}	0.060	0.064	0.666	694.2
iUNA	label, w_1	0.122	0.013	0.652	236.8
	label, w_1^S	0.095	0.020	0.639	251.8
	label, w_1^D	0.106	0.057	0.661	242.4
	label, w_1^{SD}	0.070	0.092	0.661	262.2
	label, w_2	0.136	0.028	0.647	235.0
	label, w_2^S	0.120	0.026	0.649	228.4
	label, w_2^D	**0.143**	0.035	0.661	200.6
	label, w_2^{SD}	0.117	0.070	0.664	295.6
	comment, w_1	0.097	0.002	0.636	141.2
	comment, w_1^S	0.106	0.011	0.626	126.2
	comment, w_1^D	0.123	0.046	0.639	193.0
	comment, w_1^{SD}	0.120	0.054	0.631	134.8
	comment, w_2	0.117	0.003	0.639	173.8
	comment, w_2^S	0.086	0.014	0.634	192.2
	comment, w_2^D	0.127	0.033	0.640	166.0
	comment, w_2^{SD}	0.109	0.057	0.637	191.2

Best performance without S,D 0.122 0.671

ADAPTING OUR APPROACH FOR DOMAIN-SPECIFIC PROBLEMS

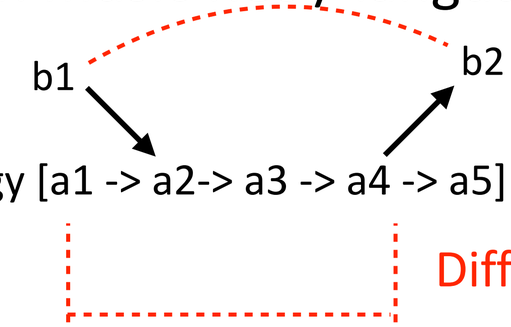


Civil registries (18th century Dutch data):

- If two people are from the same family, born in different years, have different parents, they should not be identical.

[coming soon @ ESHD'23 Conference]

Concept drifting in music KGs / language evolution

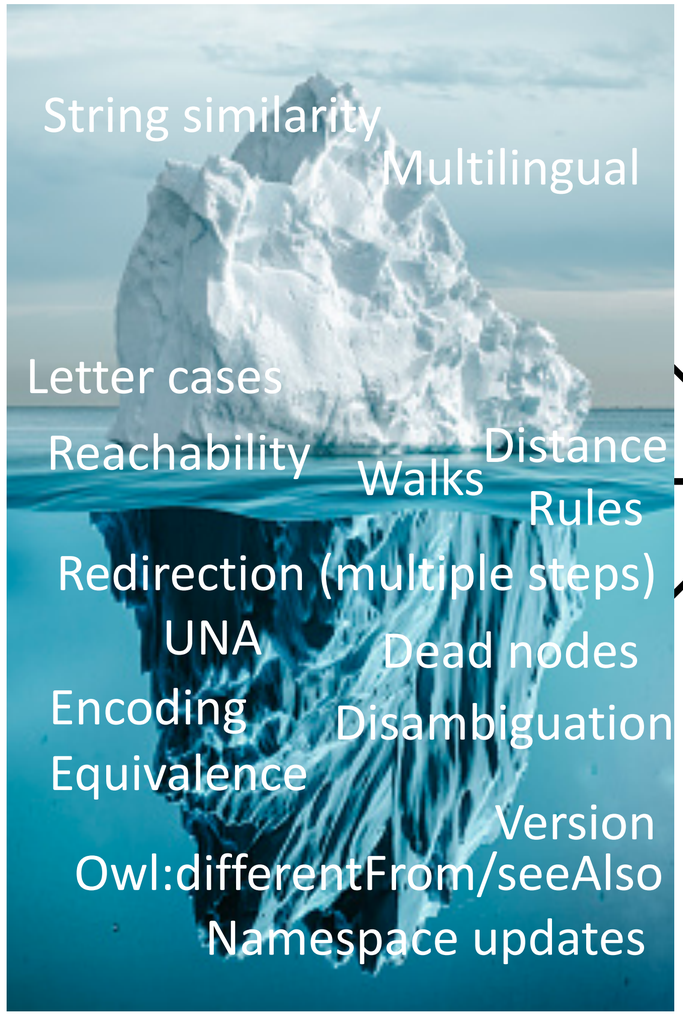


- Music Meta Ontology [a1 -> a2-> a3 -> a4 -> a5]
- Much smaller CCs

Different enough

[Idea for future collaboration with U'Bologna]

And more…



GCN, RDF2Vec

Thank you

Open source code: https://github.com/shuaiwangvu/sameAs-iUNA
Data: https://zenodo.org/record/7765113
DOI: 10.5281/zenodo.7765113

Reach out for more details and discussion: shuai.wang@vu.nl

université PARIS-SACLAY

Stay tuned for the new sameas.cc !

VU

Backup slides

Refining the results

- many singletons left after refinement

- lack of understanding of context

- move their neighbours

- link the singletons back to the graph

		#singletons	PS (%)	PU (%)	PT (%)
Louvain	res=0.01	4947.0	84.93	13.03	2.05
	res=1.0	4499.0	81.62	11.22	71.57
MetaLink	t=0.9	127	41.73	15.75	42.51
	t=0.99	57	17.54	12.28	70.17
nUNA	label, w1	486.2	79.94	15.35	4.71
	label, w2	473.2	80.40	17.09	2.51
	comment, w1	112.2	95.74	3.97	0.29
	comment, w2	103.2	94.27	4.10	1.64
qUNA	w1	226.4	54.73	43.07	2.20
	w2	202.6	43.44	54.62	1.93
iUNA	label, w1	116.4	41.27	54.24	4.50
	label, w2	111.6	32.60	62.92	4.49
	comment, w1	33.2	82.96	1.75	15.29
	comment, w2	32.0	87.92	5.39	6.69

Table 6: Singletons and their semantics.

Backup slides

Weight distribution

Weighted sameas.cc graph

Are those removed links with lower weights?

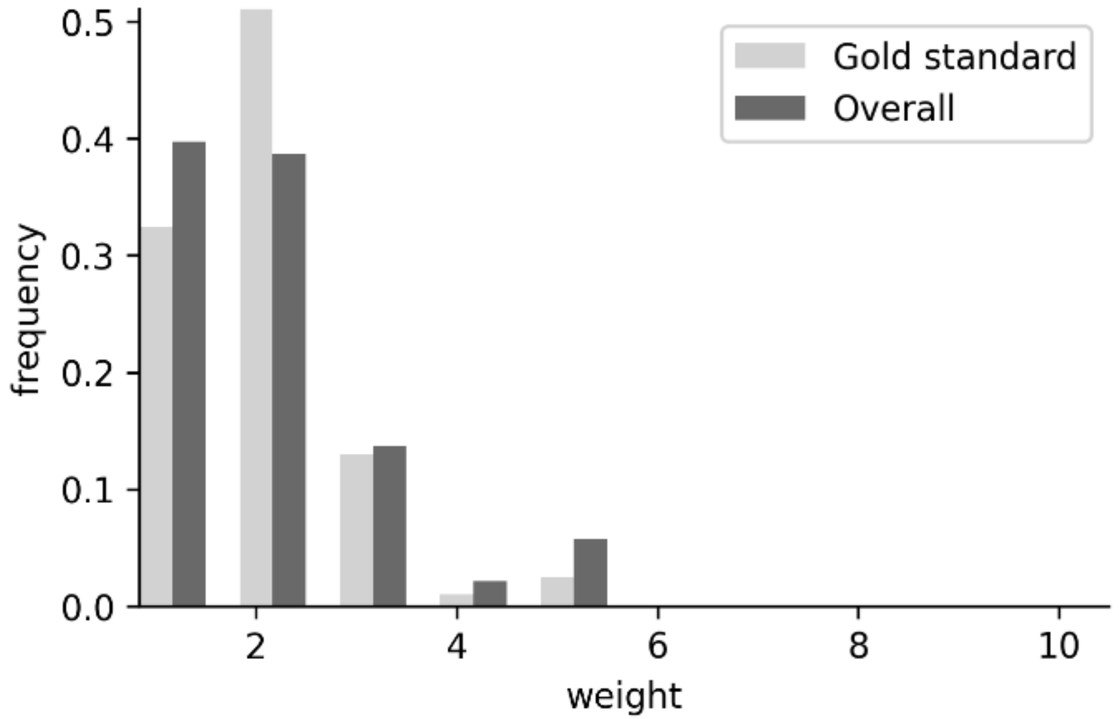


Fig. 4: Weight distribution of the `owl:sameAs` links in the LOD Laundromat.

Source of error

Ambiguity in multilingual labels.

Automated processing of information (blindly include edges in transitive closure)

Duplication in datasets

DBpedia's disambiguation: in our gold standard, we found that among the 3,678 erroneous edges, only 5 entities have multiple label-like or comment-like sources. This indicates that redundancy is not the direct cause of the error.

Reference

[1] Aidan Hogan et al. “Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora”.

[2] Joe Raad. “Identity Management in Knowledge Graphs”. doctoral dissertation. PhD thesis. University of Paris-Saclay, 2018.

[3] Andre Valdestilhas et al. “CEDAL: Time-Efficient Detection of Erroneous Links in Large-Scale Link Repositories”.

[4] Gerard de Melo. “Not Quite the Same: Identity Constraints for the Web of Linked Data

Images in the slides:

https://www.istockphoto.com/photo/iceberg-with-above-and-underwater-gm1216823850-354965398

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.yannarthusbertrandphoto.com%2Fproduit%2Fboat-icebergs%2F&psig=AOvVaw2-JfUscXLdLyEoDHeAW0fW&ust=1685606356006000&source=images&cd=vfe&ved=0CBMQjhxqFwoTCICpjJOLn_8CFQAAAAAdAAAAABAg

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fillustrations%2Fpeople-eating&psig=AOvVaw1NK8QUhHyTlNHlzSrcvZi6&ust=1685606297350000&source=images&cd=vfe&ved=0CBMQjhxqFwoTCMDt3veKn_8CFQAAAAAdAAAAABAD