KNOWLEDGE GRAPH COMPLETION
PART 3: IDENTITY LINK VALIDATION

FATIHA SAÏS [1]

NATHALIE PERNELLE[1]

DANAI SYMEONIDOU[2]

[1] LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

[2] INRA, GAMMA TEAM

[3] DEPT. OF COMPUTER SCIENCE, VU UNIV. AMSTERDAM, NL

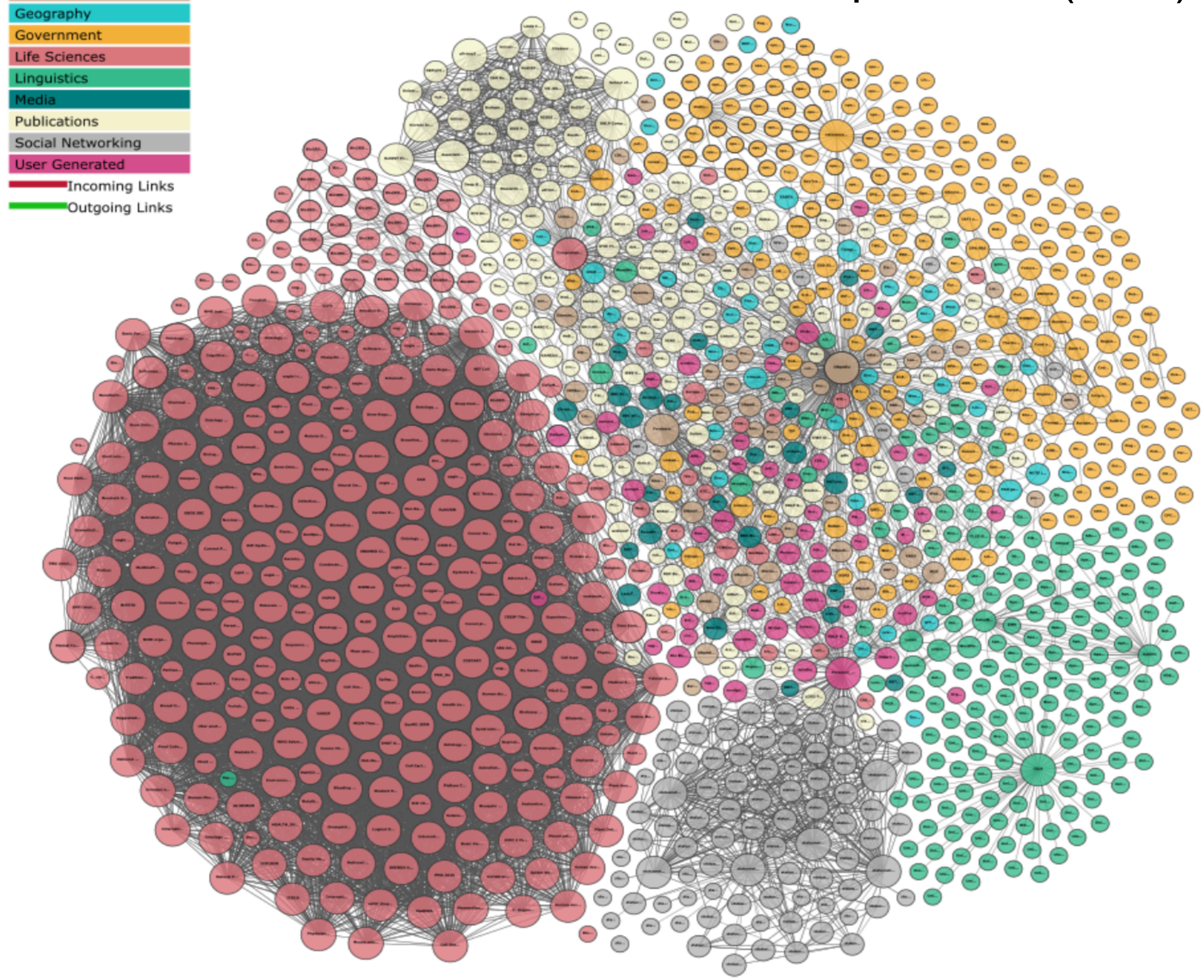
LINKED OPEN DATA

Linked Data - Datasets under an open access

- 1,139 datasets
- over 100B triples
- **about 500M links**
- several domains

Ex. DBPedia : 1.5 B triples

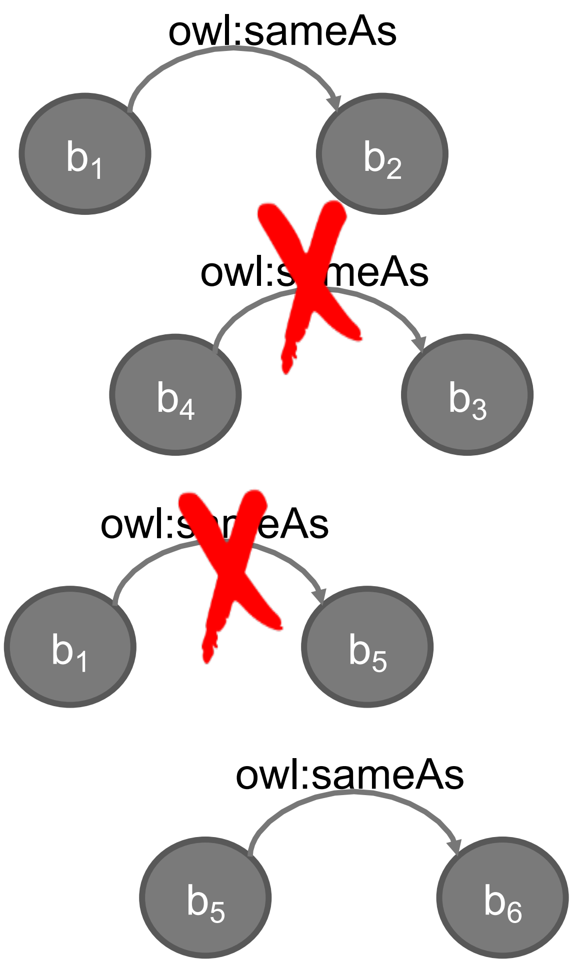
Linked Open Data (LOD)



"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

IDENTITY PROBLEM

- [Halpin et al. 2010] showed that 37% of owl:sameAs links randomly selected among 250 identity links between books were incorrect.
- In [Jaffri et al., 2008], the authors discuss how erroneous use of owl:sameAs in the interlinking of the DBpedia and DBLP datasets has resulted in publications becoming incorrectly assigned to different authors.
- Automatic data linking tools do not guarantee 100% precision, because of:
  - Errors, missing information, data freshness, etc.



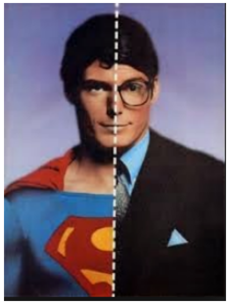
IDENTITY PROBLEM

Today, the **classical definition** of **identity** has become the **canonical** one on the **Semantic Web** (through owl:sameAs predicate).

There are some problems with it,

① Identity does not hold **across modal contexts**

- *Allow Lois Lane to believe that Superman saved her without requiring her to believe that Clark Kent saved her.*



* https://pridiyawulan.blogspot.fr/2015/01/clark-kent-atau-superman-pilih-yang-mana.html

IDENTITY PROBLEM

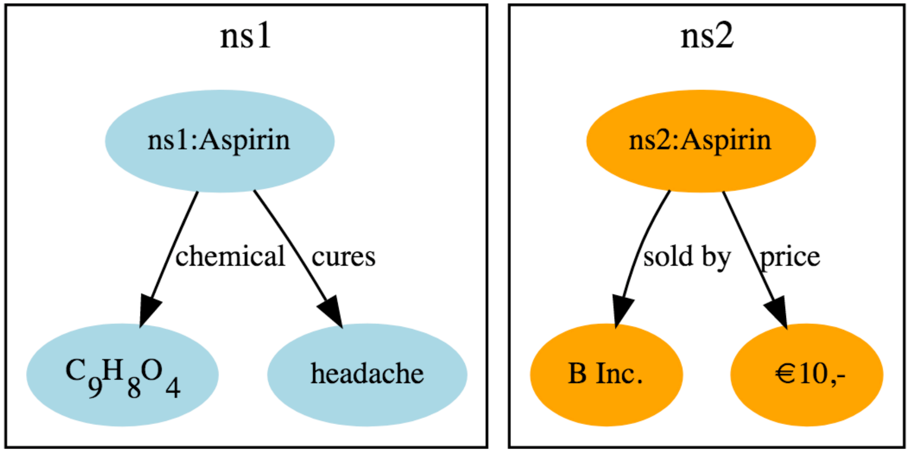
Today, the **classical definition** of **identity** has become the **canonical** one on the **Semantic Web** (through owl:sameAs predicate).

There are some problems with it,

① Identity does not hold **across modal contexts**

② Identity is **context-dependent** [Geach, 1967]

- *allowing two medicines to be considered the same in terms of their chemical substance, but different in terms of their price (e.g., because they* are produced by different companies).



*

* wouterbeek.github.io

IDENTITY PROBLEM

Today, the **classical definition** of **identity** has become the **canonical** one on the **Semantic Web** (through owl:sameAs predicate).

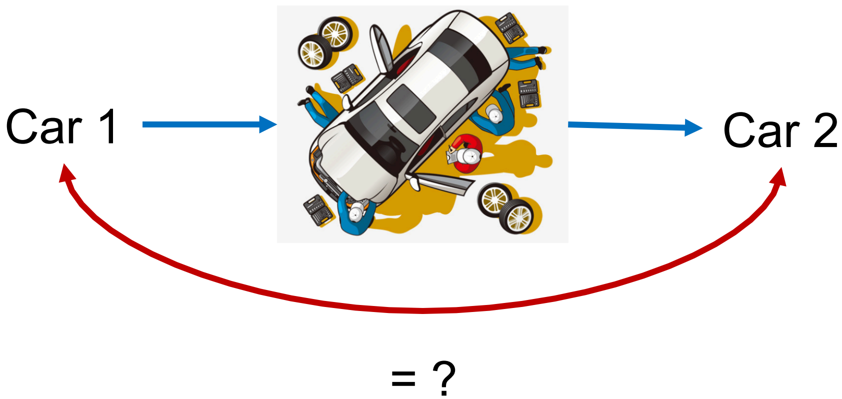
There are some problems with it,

① Identity does not hold **across modal contexts**

② Identity is **context-dependent** [Geach, 1967]

③ **Identity over time** poses problems

- since a car may be considered the same car, even though some (or even all) of its original components have been replaced by new ones.



OWL:SAMEAS PREDICATE

- owl:sameAs, indicates that two different descriptions refer to the same entity
- a strict semantics,
  1) Reflexive,
  2) Symmetric,
  3) Transitive and
  4) Fulfils property sharing:

$$\forall X \forall Y \text{ owl:sameAs}(X, Y) \wedge p(X, Z) \Rightarrow p(Y, Z)$$

IDENTITY PROBLEM: LITERATURE REVIEW

1. Detection of erroneous identity links
2. Use of alternate links
3. Detection of contextual identity links

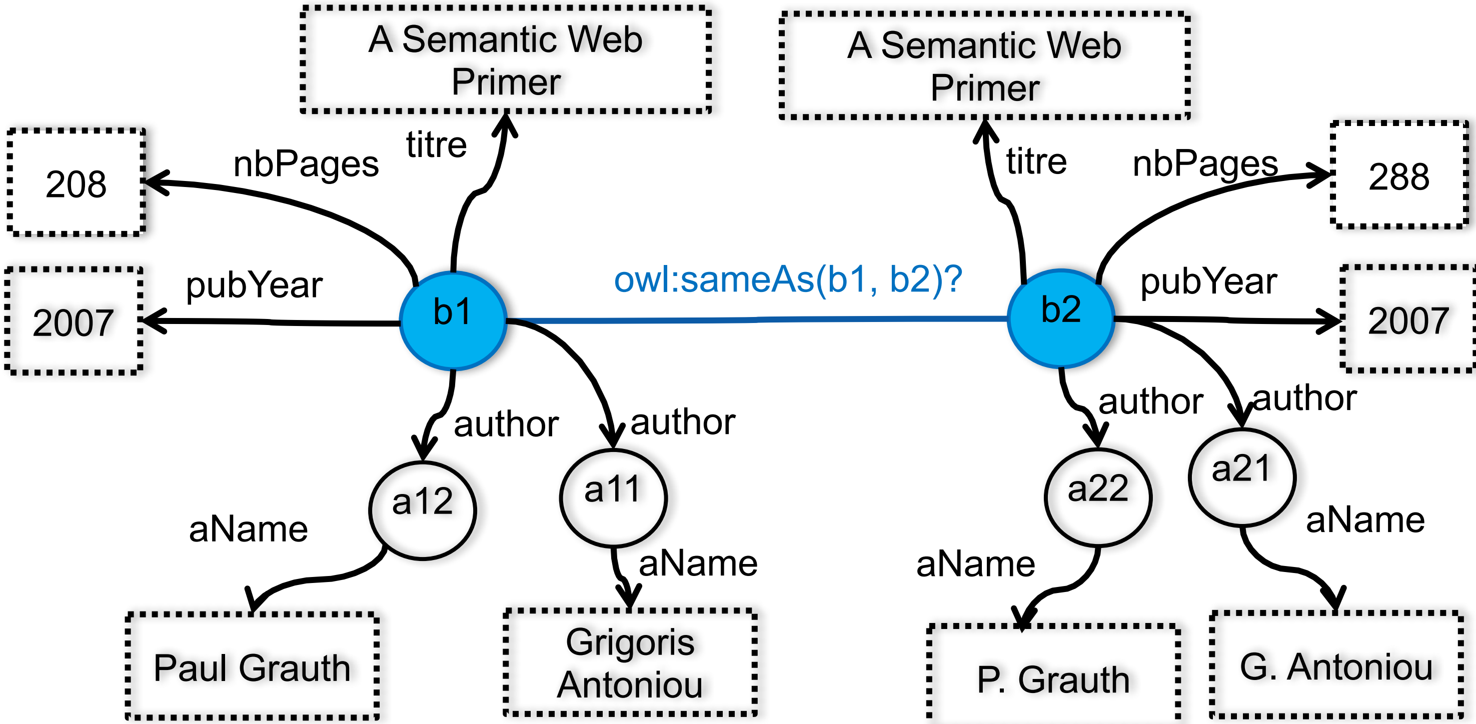
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

1. DETECTION OF ERRONEOUS IDENTITY LINKS

Content

Which kind of information to use for detecting erroneous Identity links?

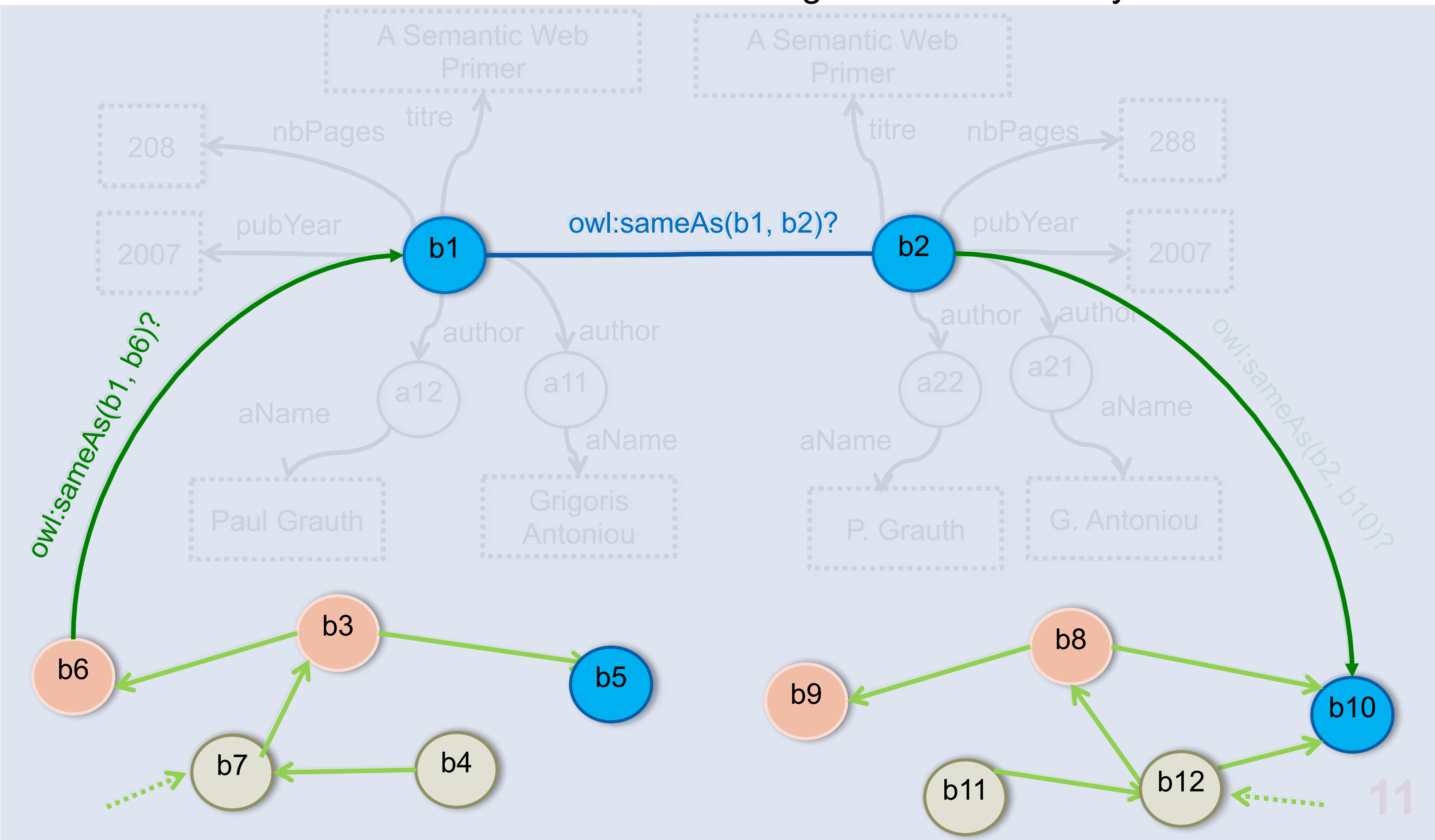


1. DETECTION OF ERRONEOUS IDENTITY LINKS

Content

Identity Network

Which kind of information to use for detecting erroneous Identity links?



1. DETECTION OF ERRONEOUS IDENTITY LINKS

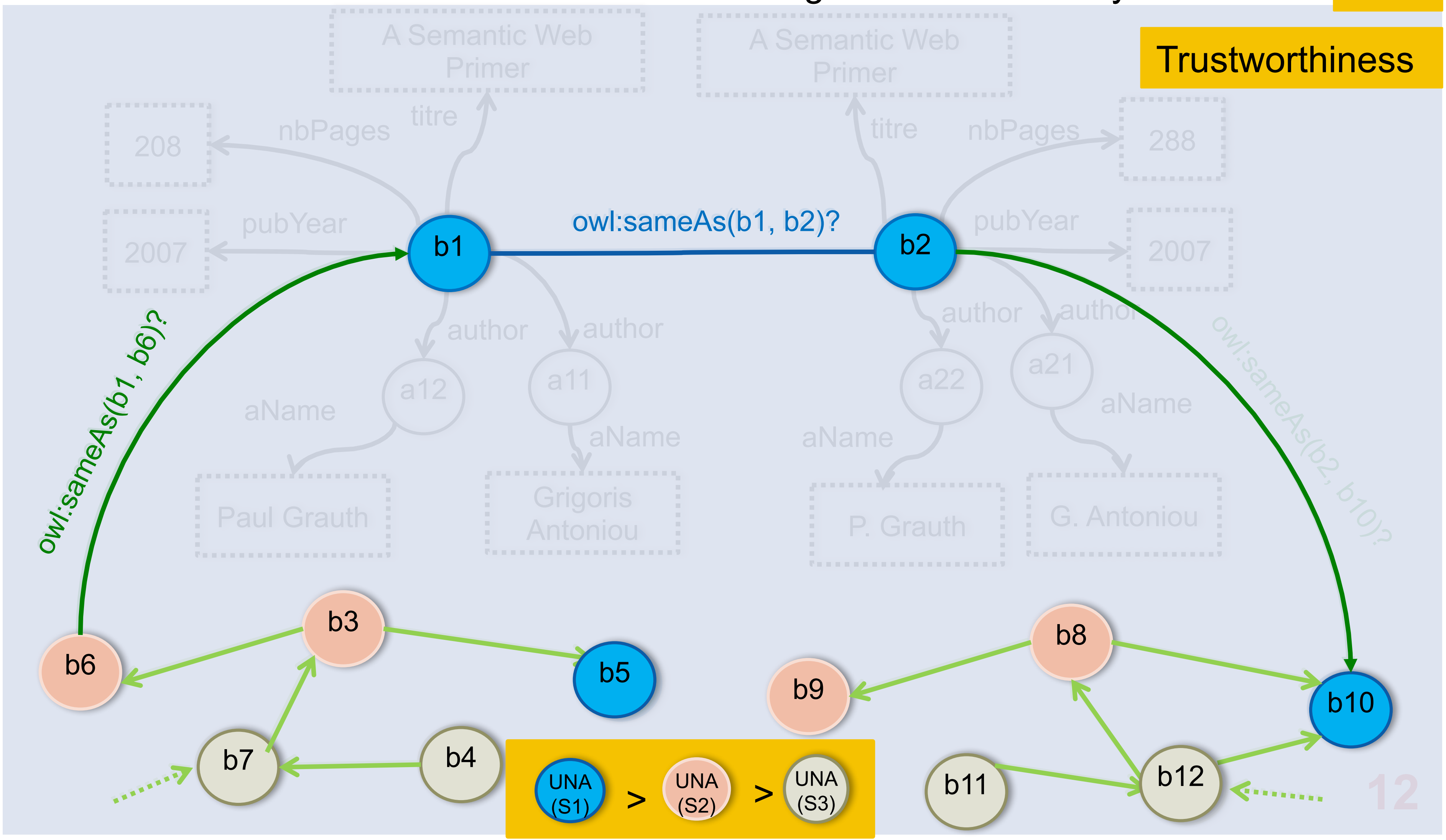
Content

Identity Network

UNA

Trustworthiness

Which kind of information to use for detecting erroneous Identity links?



1. DETECTION OF ERRONEOUS IDENTITY LINKS

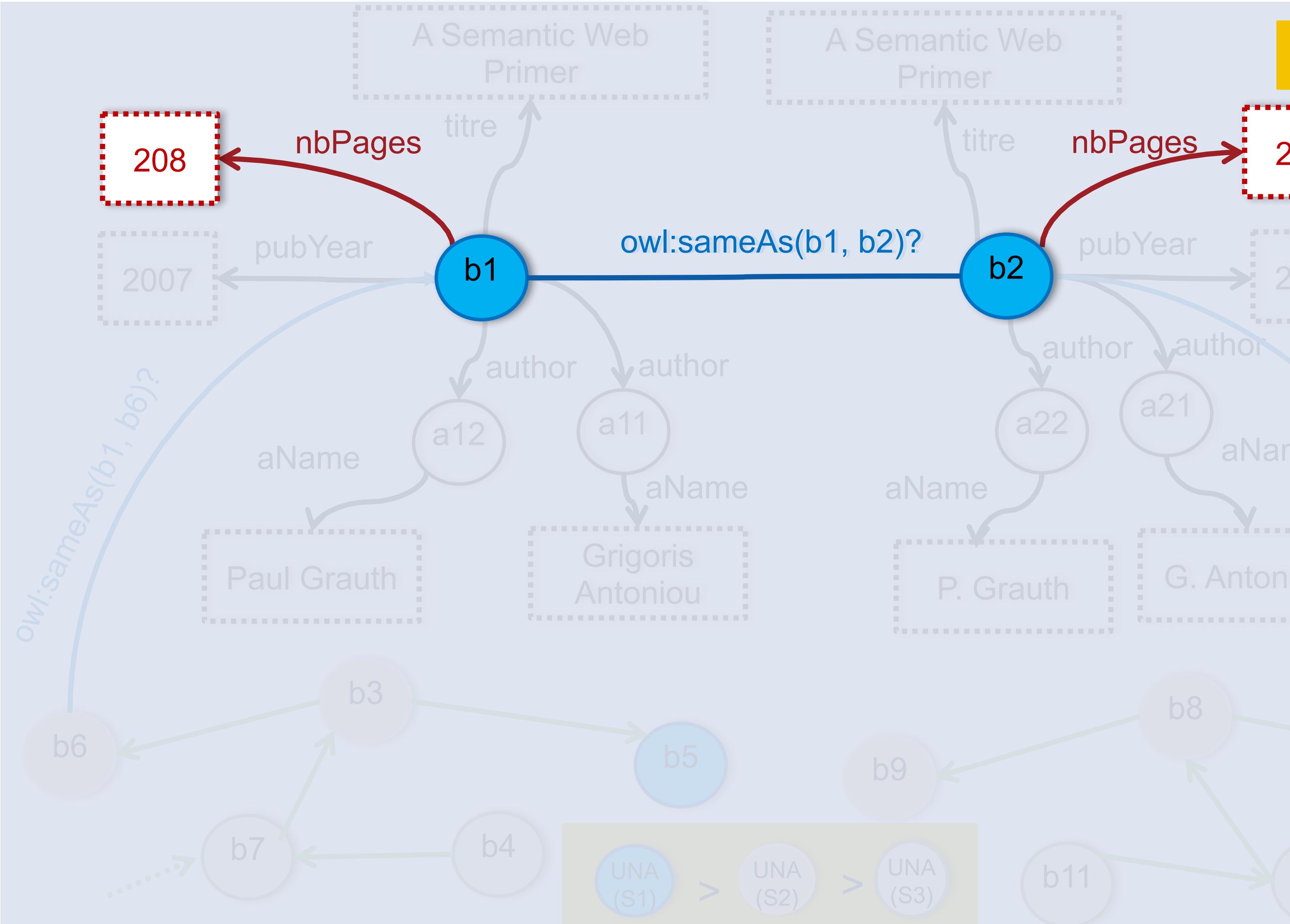
Content

Identity Network

UNA

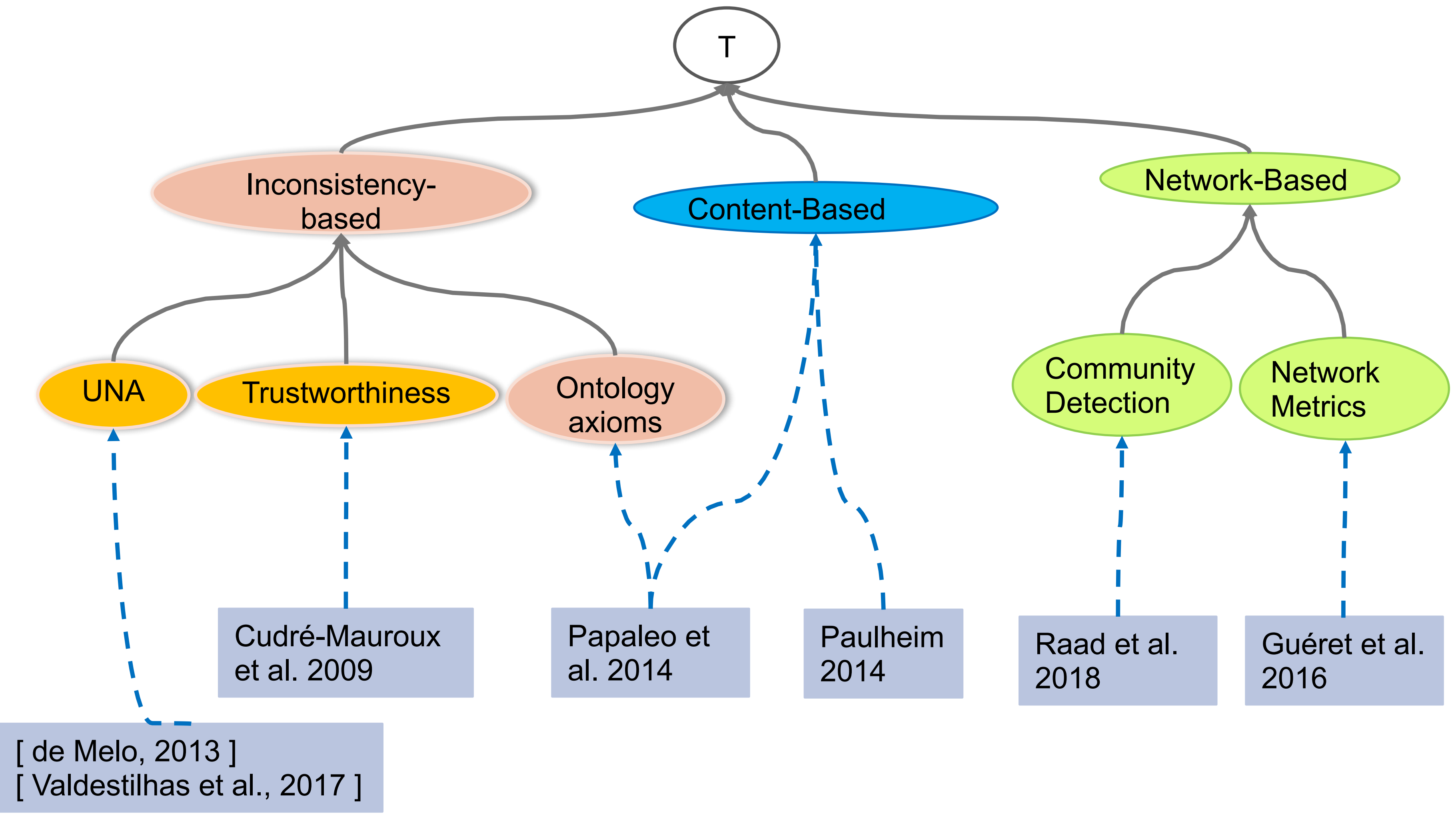
Trustworthiness

Which kind of information to use for detecting erroneous Identity links?



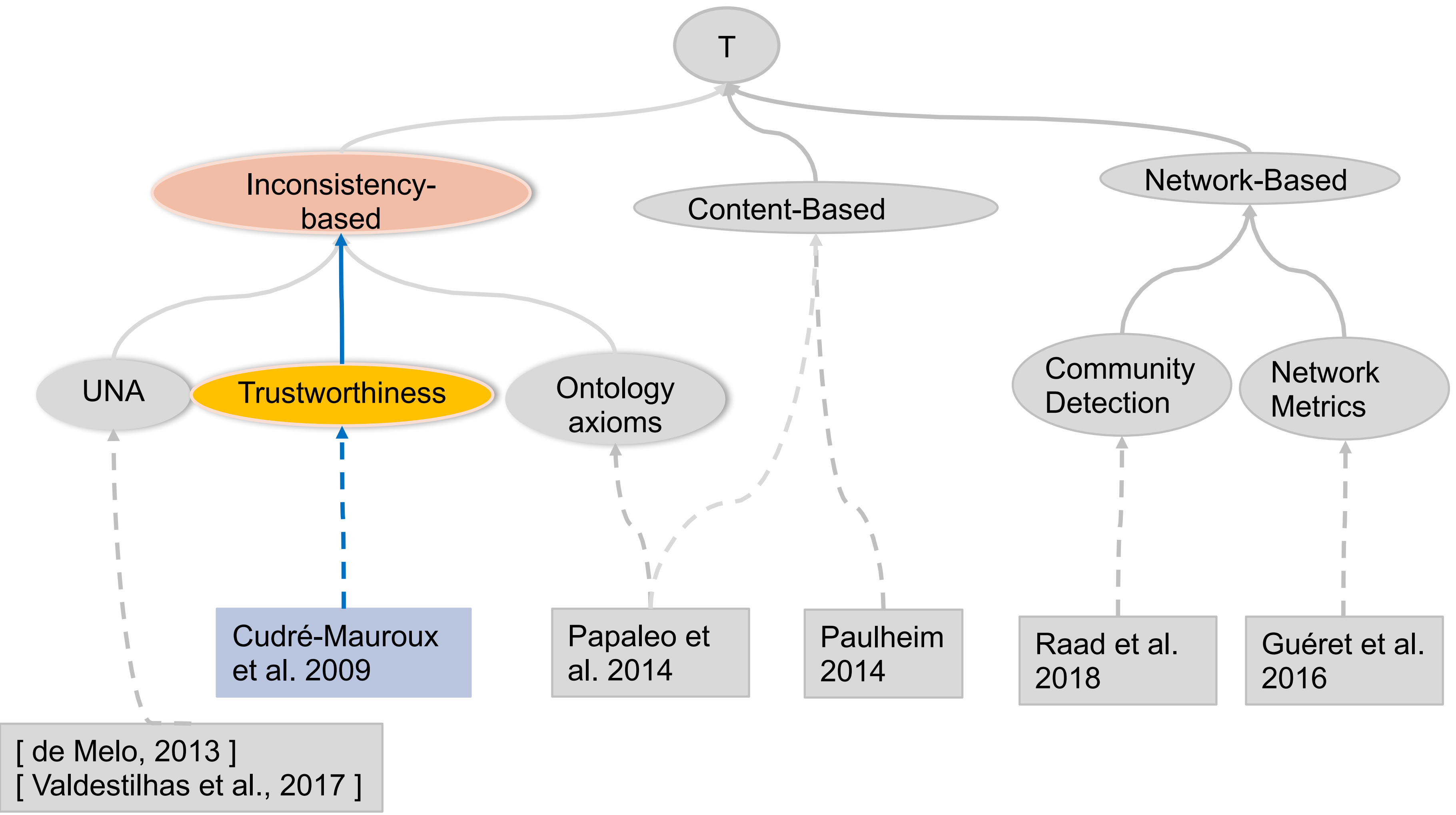
Ontology Axioms:
Func(nbPages)
LC(author)
Func(title)
Disj(Science-fiction, Memoir),
…

1. DETECTION OF ERRONEOUS IDENTITY LINKS



INCONSISTENCY-BASED

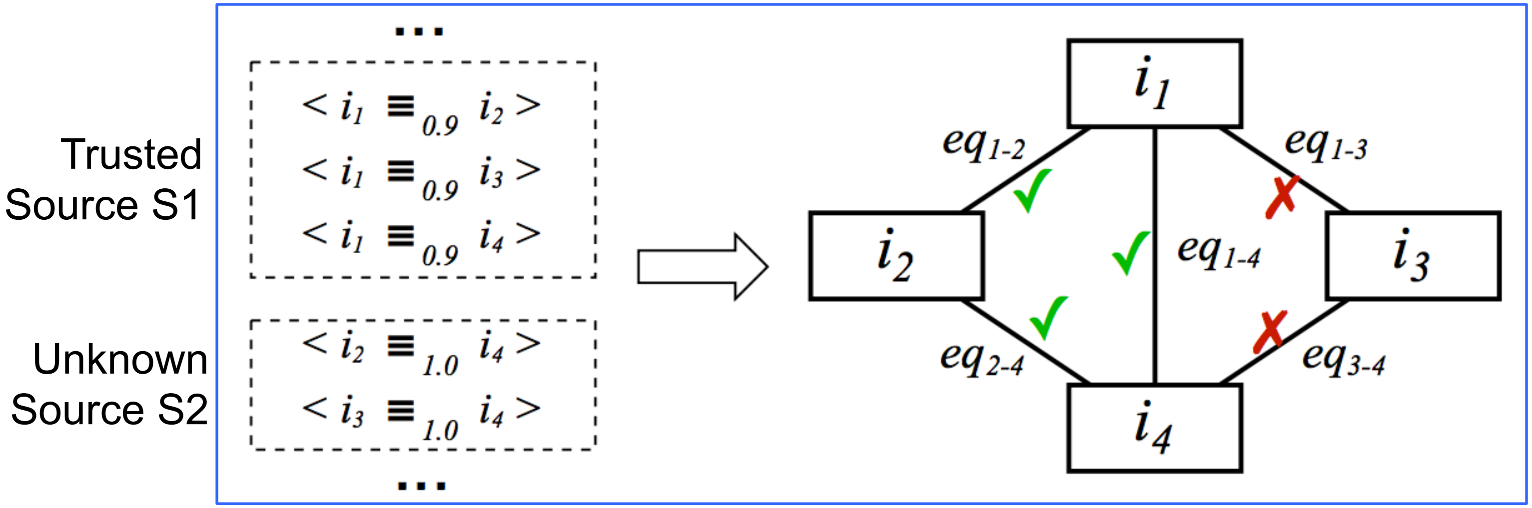
1. DETECTION OF ERRONEOUS IDENTITY LINKS



SOURCE TRUSTWORTHINESS

Cudré-Mauroux et al. 2009

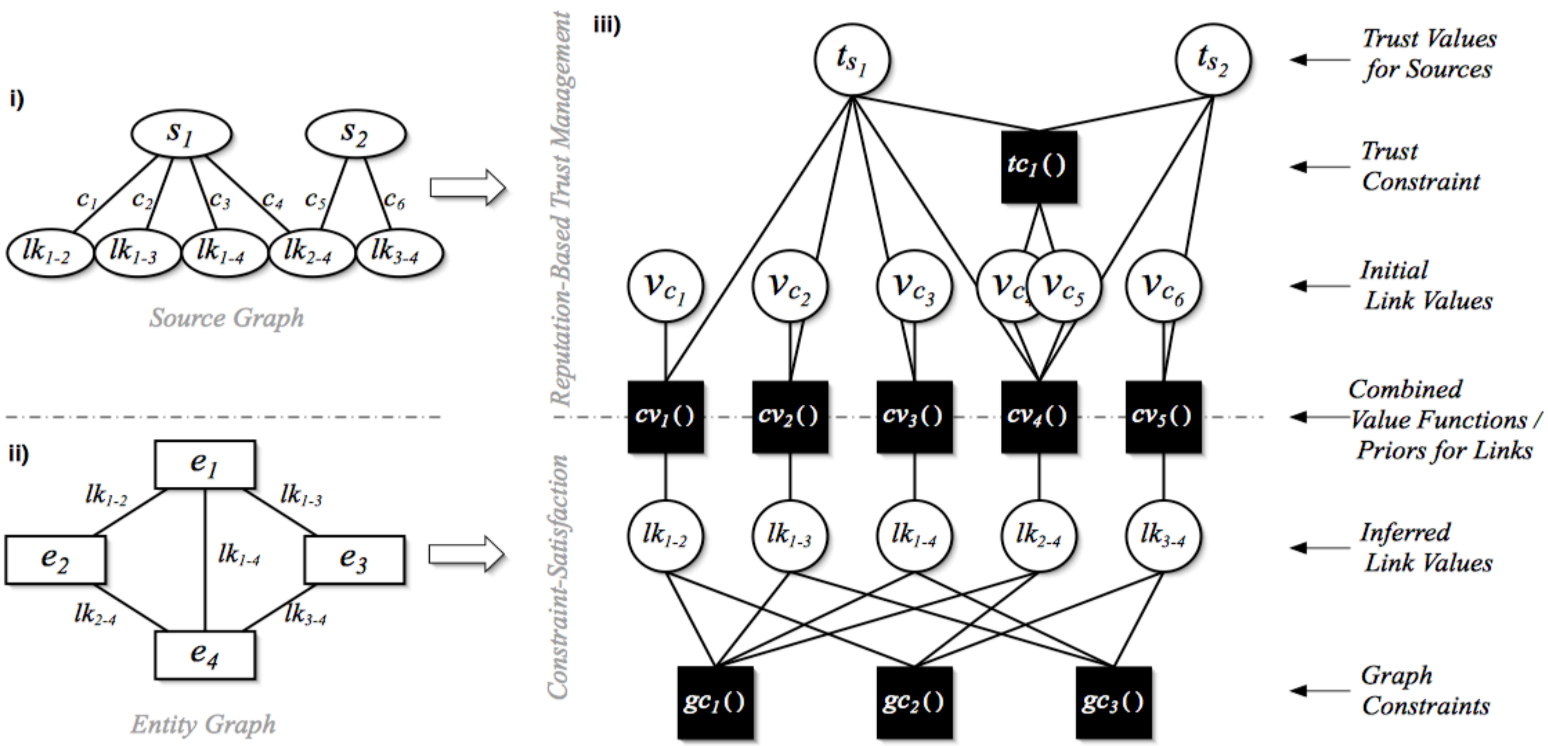
- **Principle**: owl:sameAs links published by trusted sources are more likely to be correct. Every pair of URIs coming from the same source are necessary different.
- **idMech:** a probabilistic and decentralized framework for **entity disambiguation.**



SOURCE TRUSTWORTHINESS

Probabilistic Disambiguation

Cudré-Mauroux et al. 2009



1) A graph-based constraint satisfaction problem that exploits owl:sameAs symmetry and transitivity.
2) Use of iteratively refined trustworthiness of the sources declaring the statements.

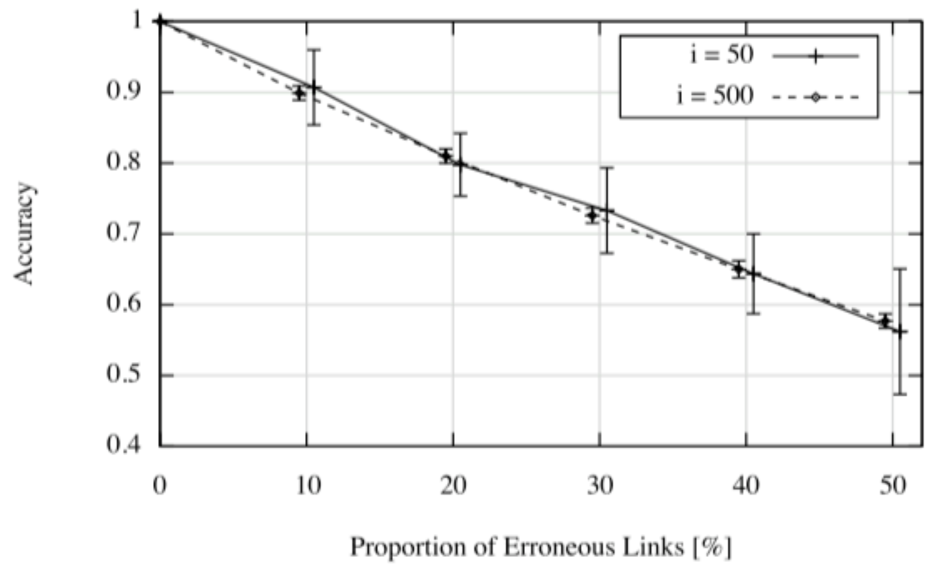
SOURCE TRUSTWORTHINESS

Evaluation on Synthetic Data

Cudré-Mauroux et al. 2009

Dataset

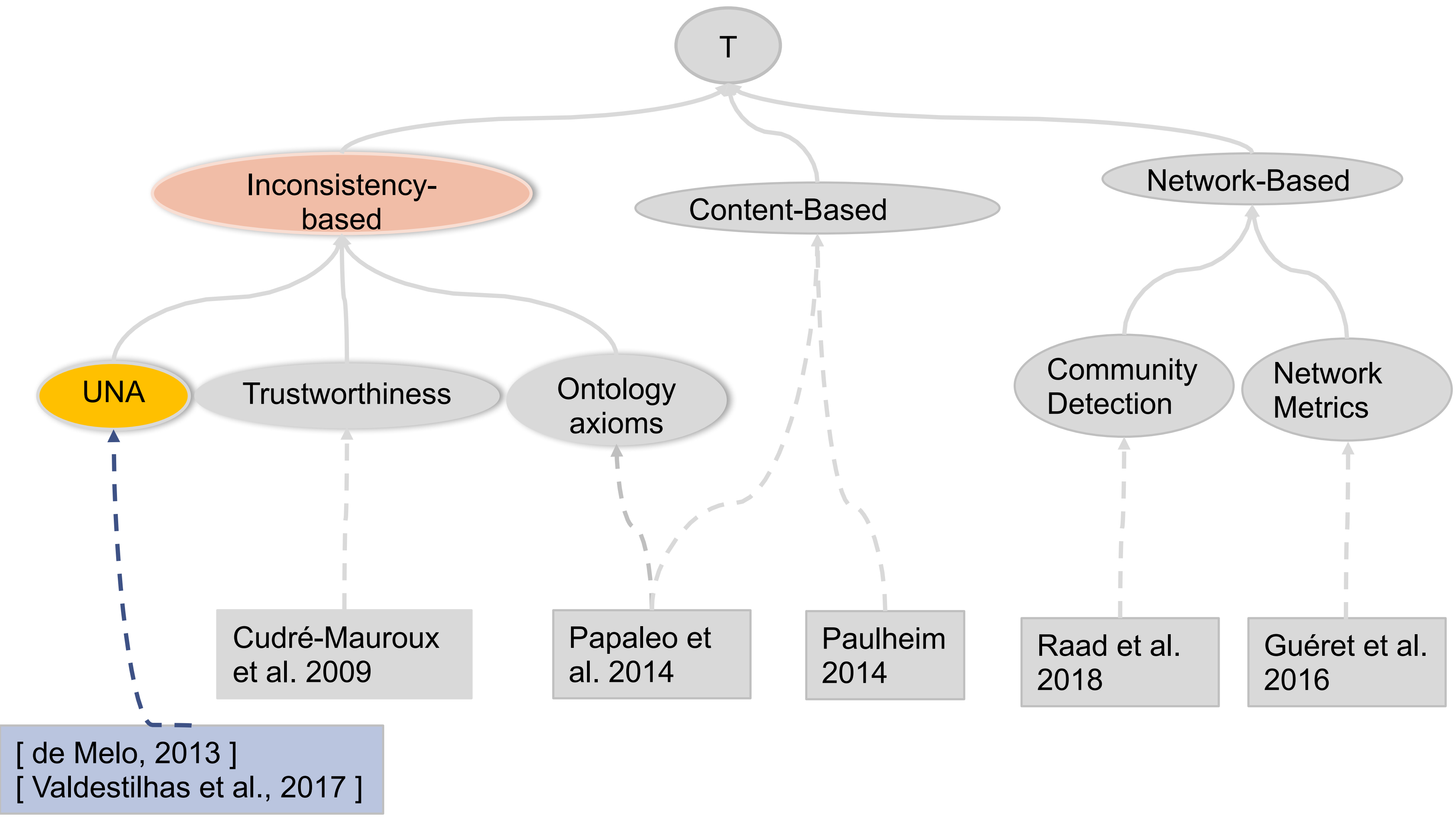
- **N**etworks of 50 and 500 entities, 150/3000 links, and a varying fraction of erroneous links (from 0 to 50%)



Results

- When considering relatively **dense networks, cycles up to size 4** and a varying fraction of **erroneous links from 0 to 50%:**
  - The more erroneous links, the lower the accuracy is.
  - The size of the graph has no impact on the accuracy of the inference.

1. DETECTION OF ERRONEOUS IDENTITY LINKS



UNIQUE NAME ASSUMPTION VIOLATION

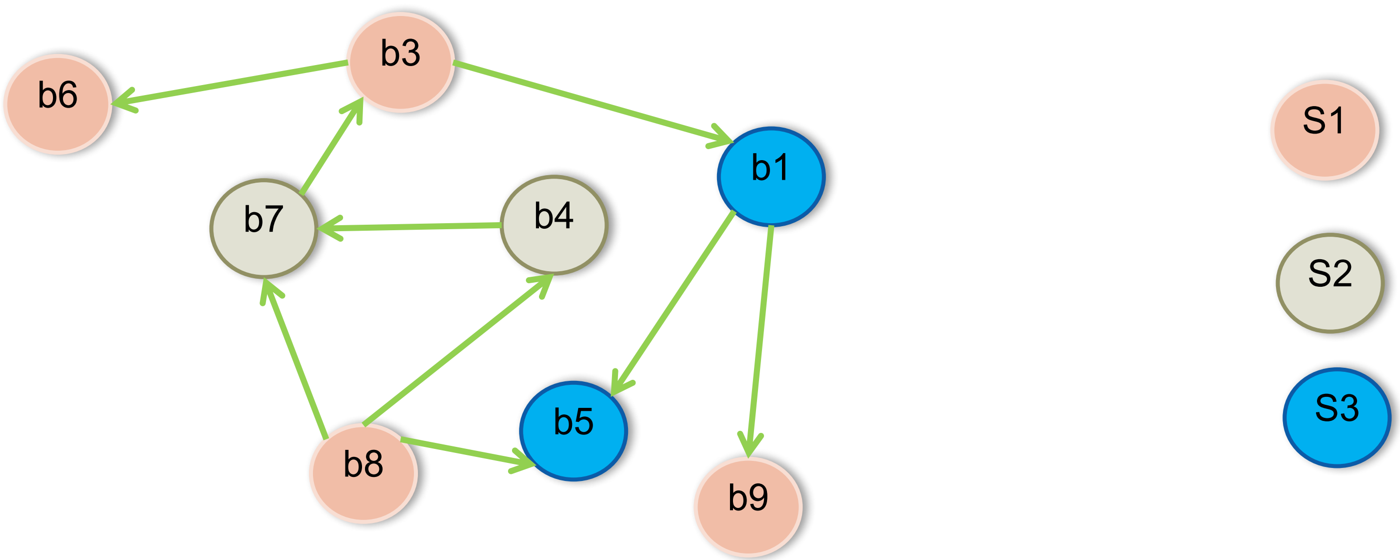
[de Melo 2013, Valdestilhas et al., 2017]

Principle

- Detecting erroneous owl:sameAs links based on Unique Name Assumption (UNA).
- The violation of the UNA is indicative of erroneous identity links.

UNA allows to state that

1. Every pair of URIs coming from the same source are necessary different.



UNIQUE NAME ASSUMPTION VIOLATION

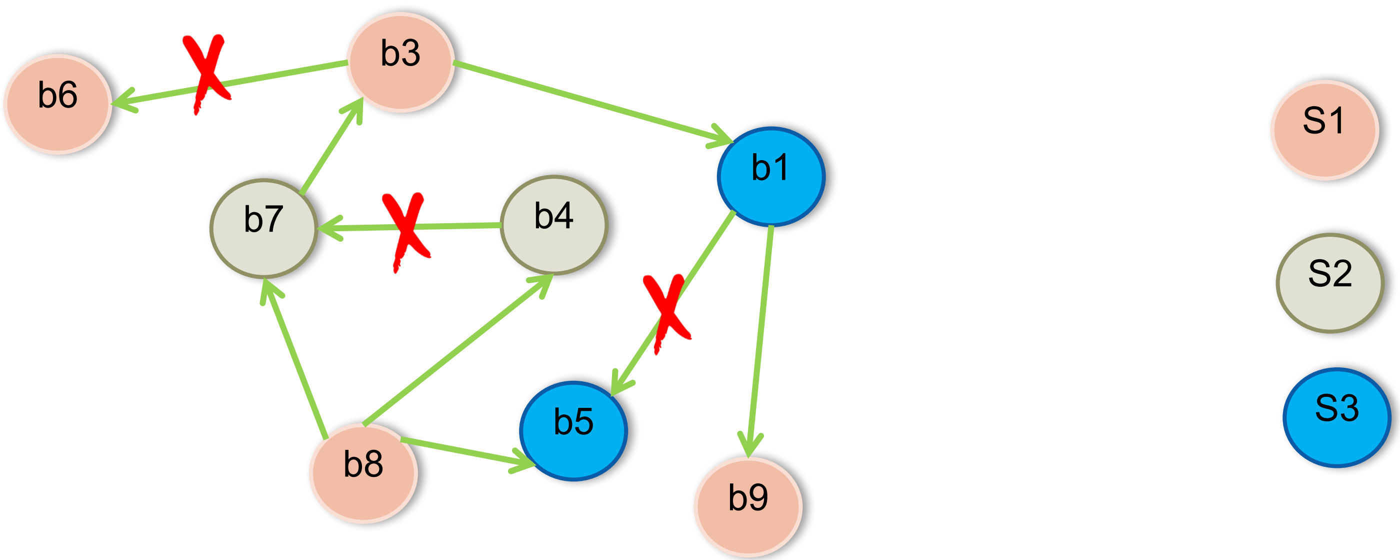
[de Melo 2013, Valdestilhas et al., 2017]

Principle

- Detecting erroneous owl:sameAs links based on Unique Name Assumption (UNA).
- The violation of the UNA is indicative of erroneous identity links.

UNA allows to state that:

1. Every pair of URIs coming from the same source are necessary different.



UNIQUE NAME ASSUMPTION VIOLATION

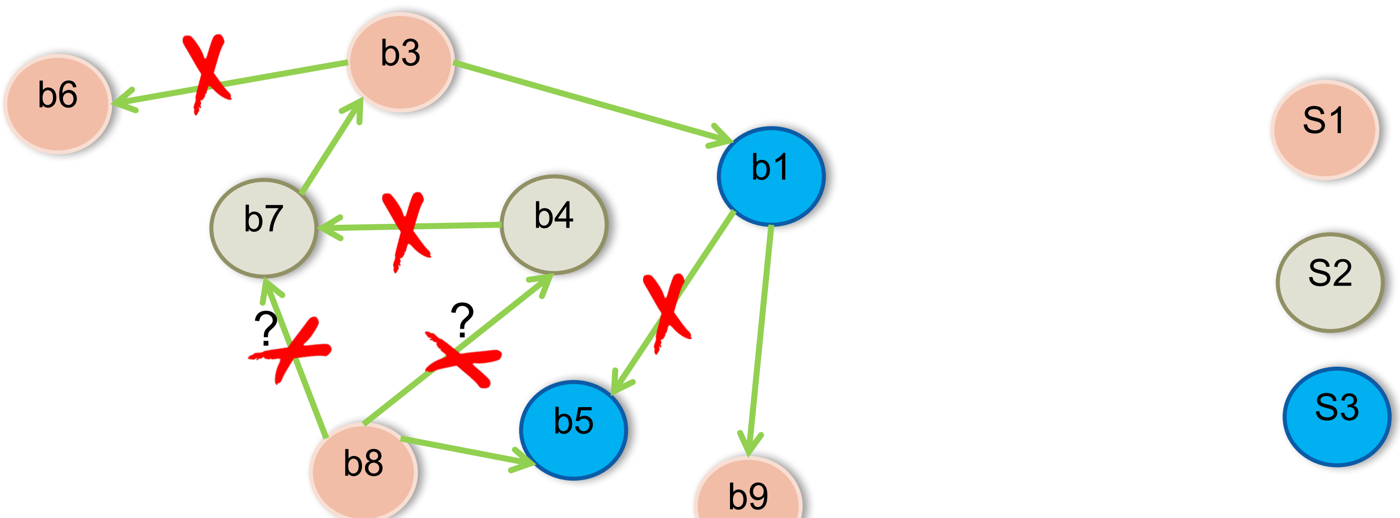
[de Melo 2013, Valdestilhas et al., 2017]

Principle

- Detecting erroneous owl:sameAs links based on Unique Name Assumption (UNA).
- The violation of the UNA is indicative of erroneous identity links.

UNA allows to state that:

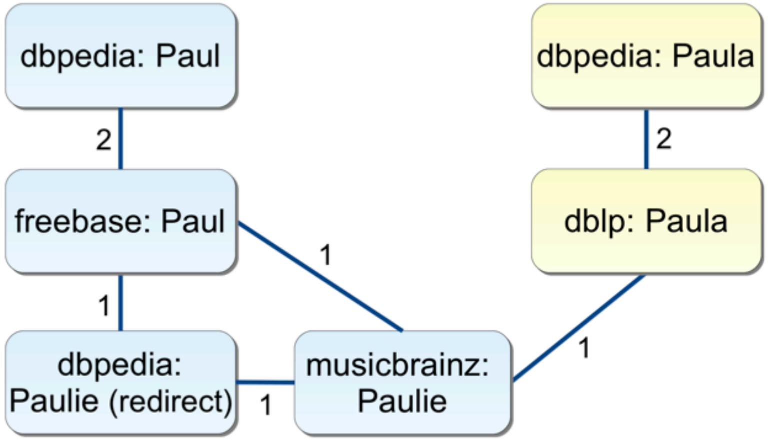
1. Every pair of URIs coming from the same source are necessary different.
2. Each URI of a source S1 cannot be identical to more than one URI of a source S2.



UNA VIOLATION

[de Melo 2013]

- Creates **undirected labeled graphs** from the existing owl:sameAs links.
- Considers a set of **distinctness constraints** to account for exceptions.



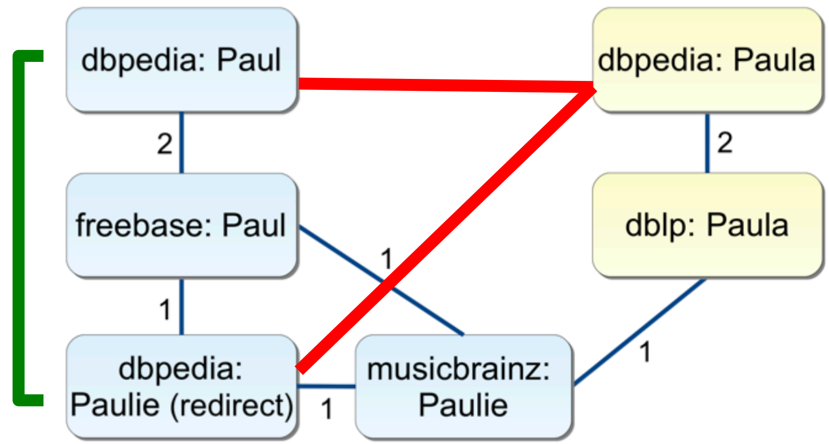
D_i({dbpedia:Paul, dbpedia:Paulie(redirect)}, {dbpedia:Paula})

UNA VIOLATION

[de Melo 2013]

- Creates **undirected labeled graphs** from the existing owl:sameAs links.
- Considers a set of **distinctness constraints** to c account for exceptions.

- Considers the problem of computing the **minimum cut** (NP-Hard Problem)
- Uses a **linear program relaxation algorithm**, that aims at deleting the minimal number of edges to cut to ensure the UNA.



D_i({dbpedia:Paul, dbpedia:Paulie(redirect)}, {dbpedia:Paula})

UNA VIOLATION

Evaluation LOD Data

[de Melo 2013]

Datasets

	#URI	Relevant Predicates			
		#sameAs	#skos:closeMatch	#skos:exactMatch	#:differentFrom
BTC2011	~4M	~3.5M	125,313	22,398	619
sameas.org 2011	~31M	22.4M			

UNA VIOLATION

Evaluation LOD Data

[de Melo 2013]

Datasets

	#URI	Relevant Predicates			
		#sameAs	#skos:closeMatch	#skos:exactMatch	#:differentFrom
BTC2011	~4M	~3.5M	125,313	22,398	619
sameas.org 2011	~31M	22.4M			

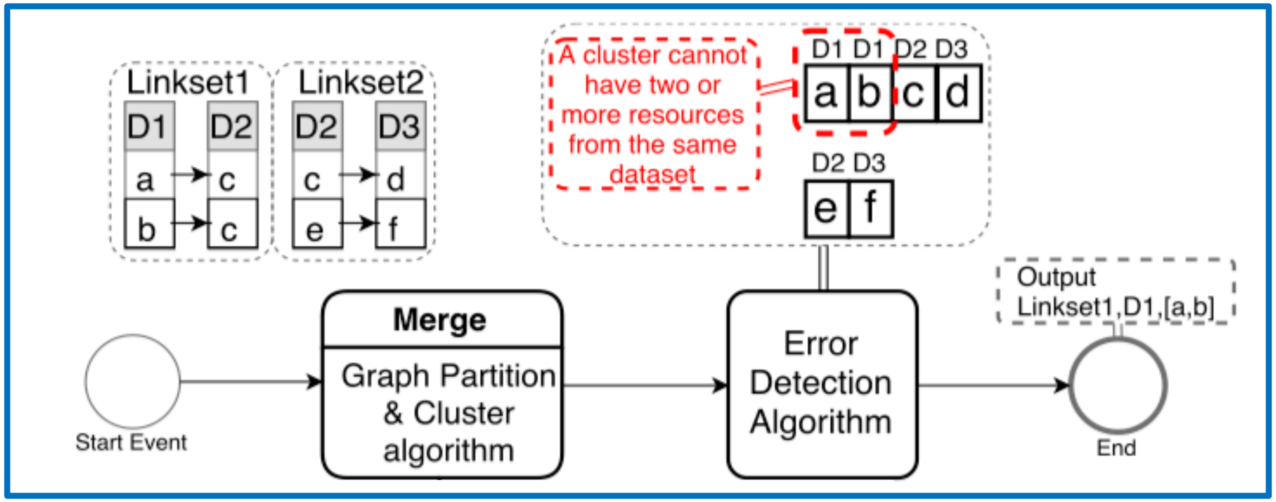
Results

- Several **hundred thousand** sameAs edges are removed automatically.
- # edges removed < # constraint violations

	BTC2011 +sameas.org	BTC2011	sameas.org
Undirected edges removed	280,086	32,753	245,987
Violations per removed edge	1.85	4.24	1.53

UNA VIOLATION

[Valdestilhas et al., 2017]



CEDAL

- **Erroneous links**: detection of resources sharing the **same equivalence** class and the **same dataset**.
- **Rate of consistent resources** inside an equivalence class

$$M1 = \frac{\sum_{P \in \mathcal{P}^-} |P|}{\sum_{P \in \mathcal{P}} |P|}$$

- **P** contains only resources belonging to the same dataset.
- **P⁻** is the set of consistent resources

- Efficient generation of equivalence classes based on *Union Find* algorithm

UNA VIOLATION

[Valdestilhas et al., 2017]

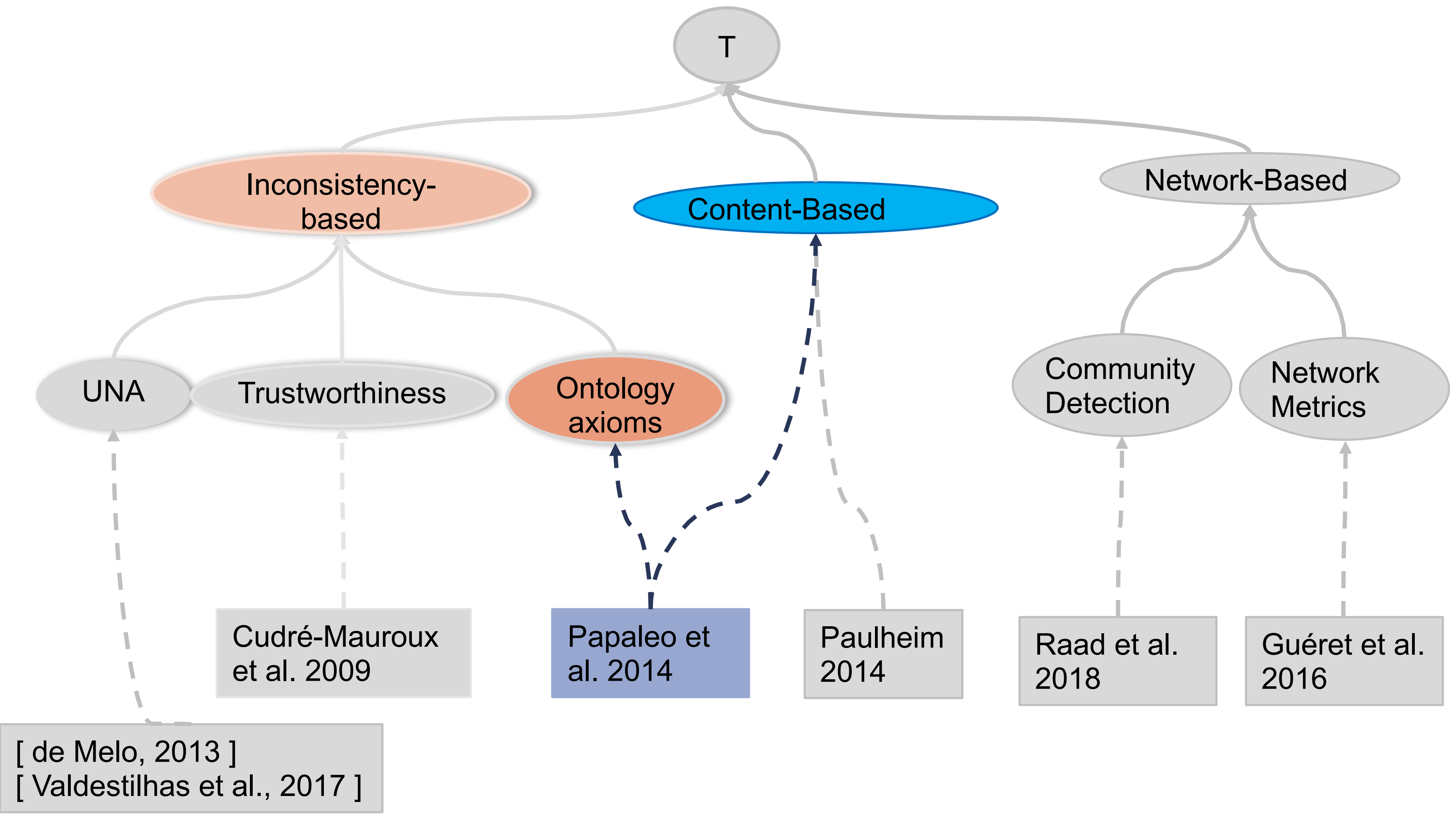
Datasets

- **LinkLion repository ~19.6M links**
- **$\mu = \frac{1}{2}(|C|(|C|-1))$,** C is the set of inconsistent resources
- K1: the knowledge base with more errors (**11.5 %**)
- K10: the knowledge base with fewer errors (**0.06 %**)

Label	Knowledge Base
K1	dotac.rkbexplorer.com—eprints.rkbexplorer.com.nt
K2	d-nb.info—viaf.org.nt
K3	dblp.rkbexplorer.com—dblp.l3s.de.nt
K4	linkedgeodata.org—sws.geonames.org.nt
K5	citeseer.rkbexplorer.com—kisti.rkbexplorer.com.nt
K6	wiki.rkbexplorer.com—oai.rkbexplorer.com.nt
K7	www4.wiwiss.fu-berlin.de—dbpedia.org.nt
K8	southampton.rkbexplorer.com—nsf.rkbexplorer.com.nt
K9	rae2001.rkbexplorer.com—newcastle.rkbexplorer.com.nt
K10	lod.geospecies.org—bio2rdf.org.nt

- **Data linking algorithms** (LIMES, SILK and DBpedia Extraction Framework) have a **better consistency index** than repositories such as sameas.org (**13%).**

1. DETECTION OF ERRONEOUS IDENTITY LINKS

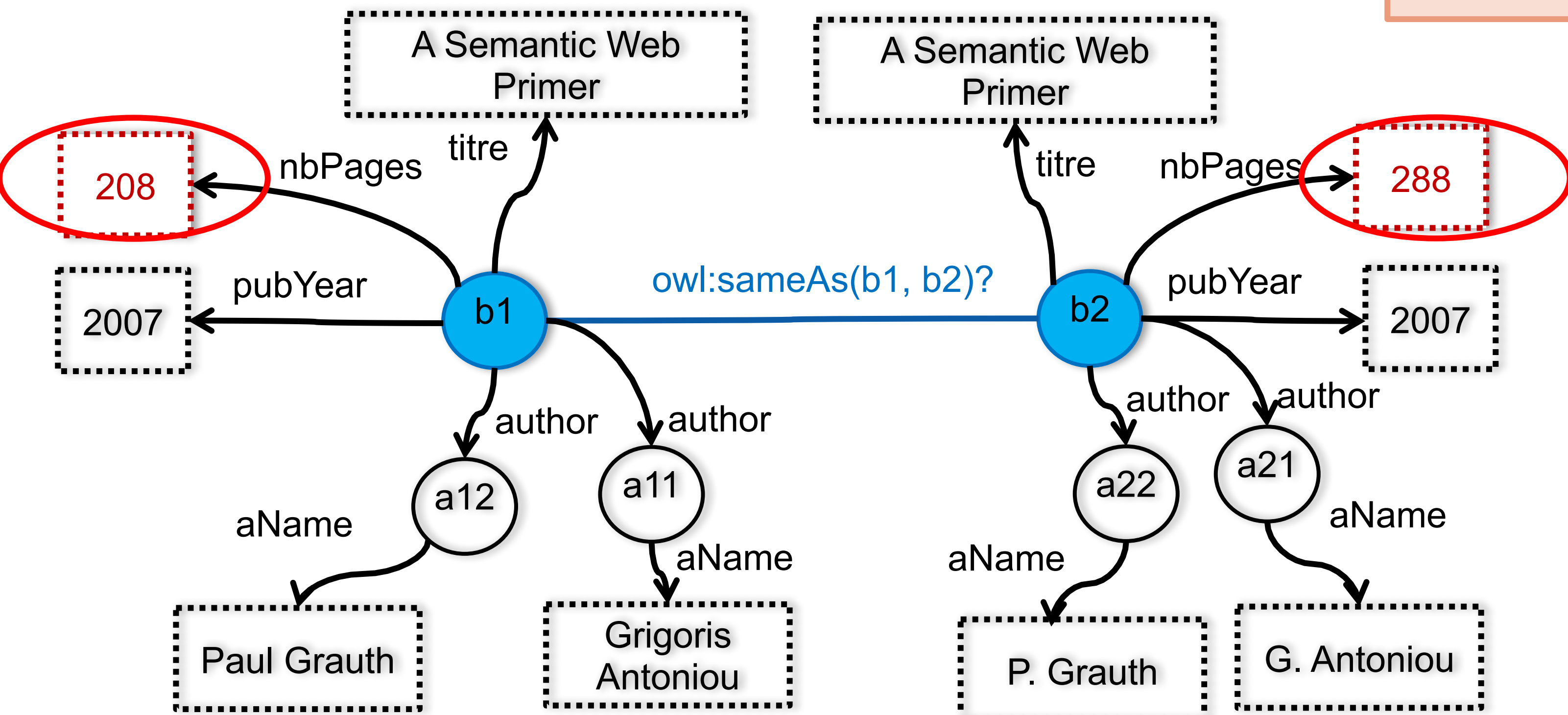


ONTOLOGY AXIOM VIOLATION

[Papaleo *et al.*, 2014]
[Hogan et al. 2012]

Principle: use of ontology axioms (functionality, local completeness, asymmetry, etc.) to detect inconsistencies or error candidates in the linked resources descriptions.

nbPages is a Functional Property



ONTOLOGY AXIOM VIOLATION

[Papaleo *et al.*, 2014]

- A logical **ontology-based method** to detect invalid sameAs statements
- Builds a contextual graph «around» each one of the two resources involved in the sameAs by exploiting ontology axioms on:
  - functionality and inverse functionality of properties and
  - local completeness of some properties, e.g., the author list of a book.
- Exploit the descriptions provided in these contextual graphs to eventually detect inconsistencies or high dissimilarities.

ONTOLOGY AXIOM VIOLATION

[Papaleo *et al.*, 2014]

Apply Unit Resolution on $\{F \cup R\}$.
[F set of facts, R set of rules]

F is the set of RDF facts
enriched by a set of ¬synVals facts in the form

$\neg synVals(w_1, w_2)$

w_1 and w_2, being literals and different.

EXAMPLES:
- notSynVals('231','100')
for a functional property *nbPages*

-notSynVals('New York', 'Paris')
for a functional property *cityName*

… knowledge from expert or extracted.

ONTOLOGY AXIOM VIOLATION

[Papaleo *et al.*, 2014]

Apply Unit Resolution on {F ∪ R}.
[F set of facts, R set of rules]

R the set of rules

(inverse) functional properties

- $R1_{FDP}$: $sameAs(x, y) \wedge p_i(x, w_1) \wedge p_i(y, w_2) \rightarrow synVals(w$
- $R2_{FOP}$: $sameAs(x, y) \wedge p_j(x, w_1) \wedge p_j(y, w_2) \rightarrow sameAs(w$
- $R3_{IFP}$: $sameAs(x, y) \wedge p_k(w_1, x) \wedge p_k(w_2, y) \rightarrow sameAs(w$

sameAs(x,y) ∧nbPages(x,w_1) ∧ nbPages(y,w_2) → SynVals(w_1,w_2)

local complete properties

- $R4_{LC}$: $sameAs(x, y) \wedge p(x, w_1) \rightarrow p(y, w_1)$

sameAs(x,y) ∧ hasAuthor(x,w_1) → hasAuthor(y,w_1)

ONTOLOGY AXIOM VIOLATION

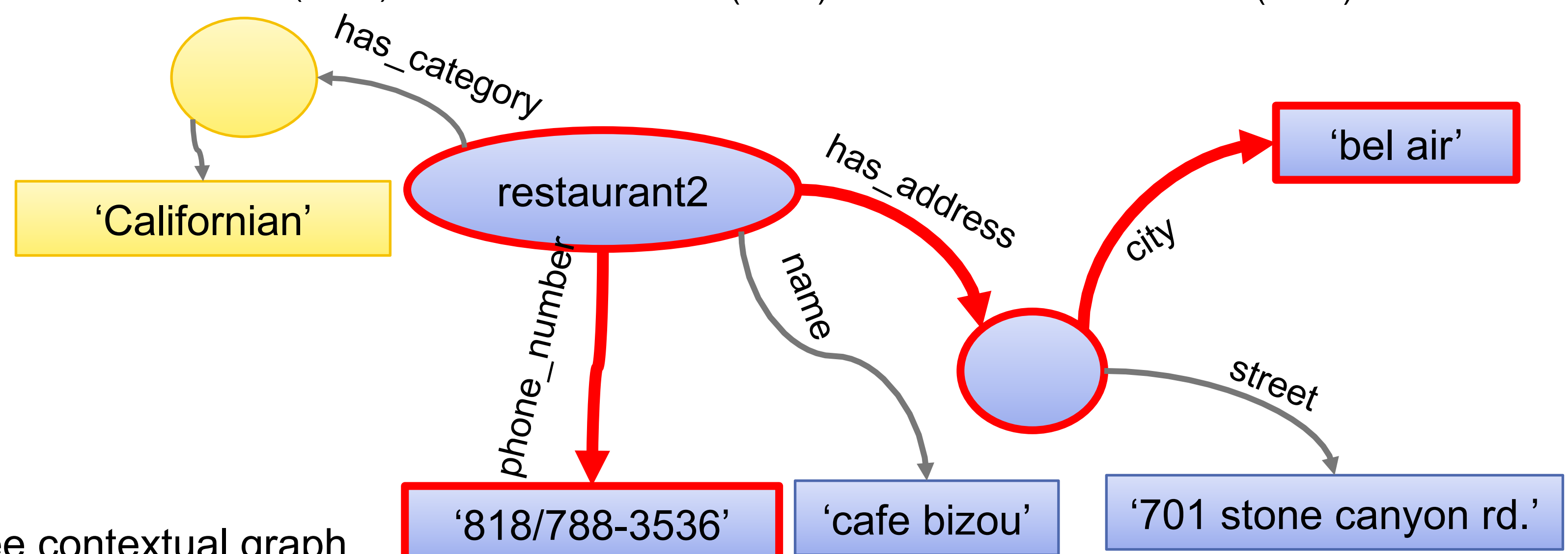
[Papaleo et al. 2014]

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].

[1] Sais et al.: *LN2R a knowledge based reference reconciliation system: OAEI2010 results*. (2010)

[2] Symeonidou et al.: *SAKey: Scalable Almost Key Discovery in RDF Data*. (2014)

[3] Yves et al.: *Ontology matching with semantic verification*. (2009)



2-degree contextual graph
phone_number, hasAddress & city
(possible synvals computation)

ONTOLOGY AXIOM VIOLATION

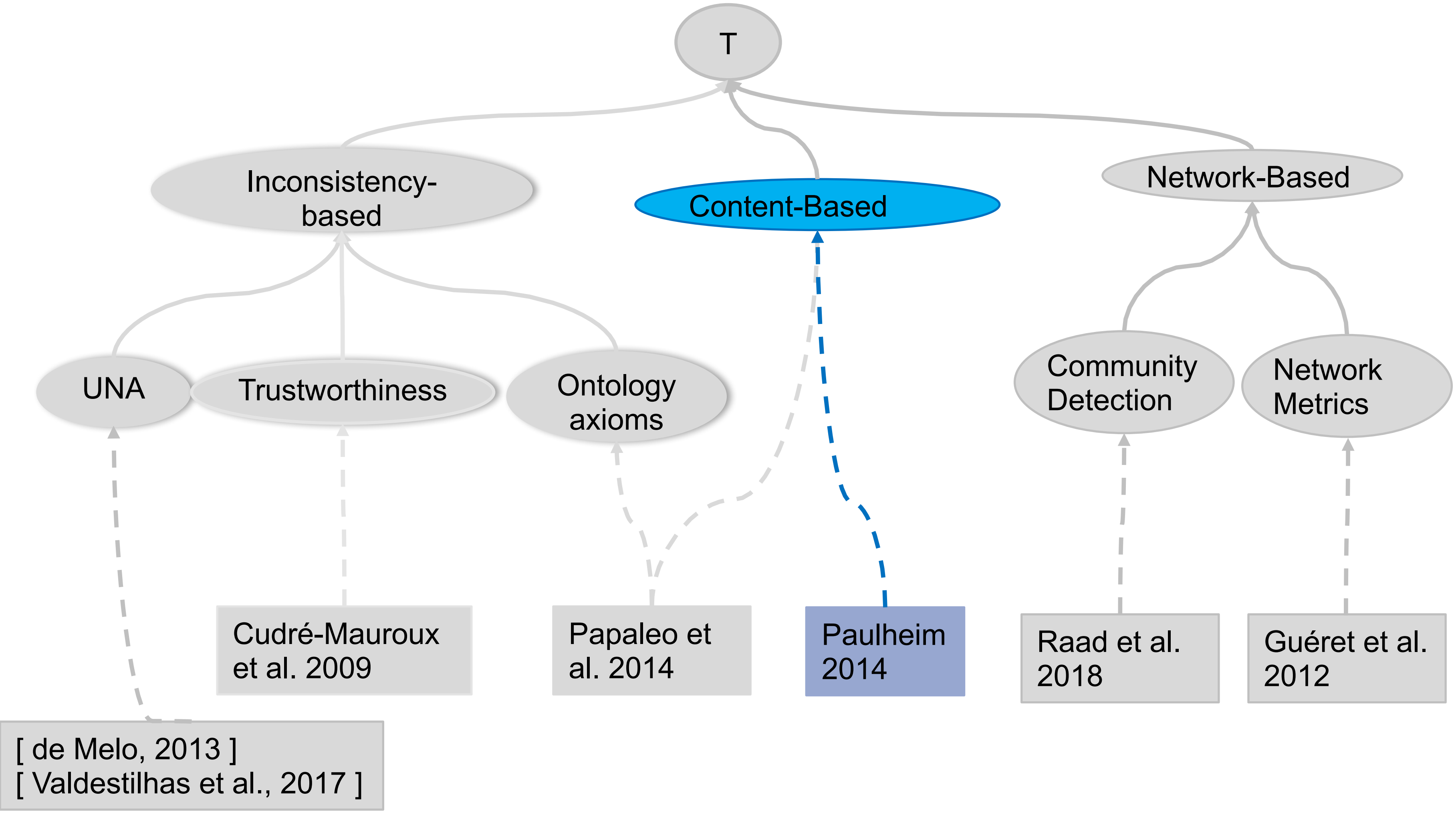
[Papaleo et al. 2014]

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].

LM	LM Precision	linkInv precision	LM+linkInv precision
2	95.55%	37%	**98.85%**
1	69.71%	88.4%	**95.19%**
3	90.17%	42.30%	**100%**

Improvement in precision

1. DETECTION OF ERRONEOUS IDENTITY LINKS

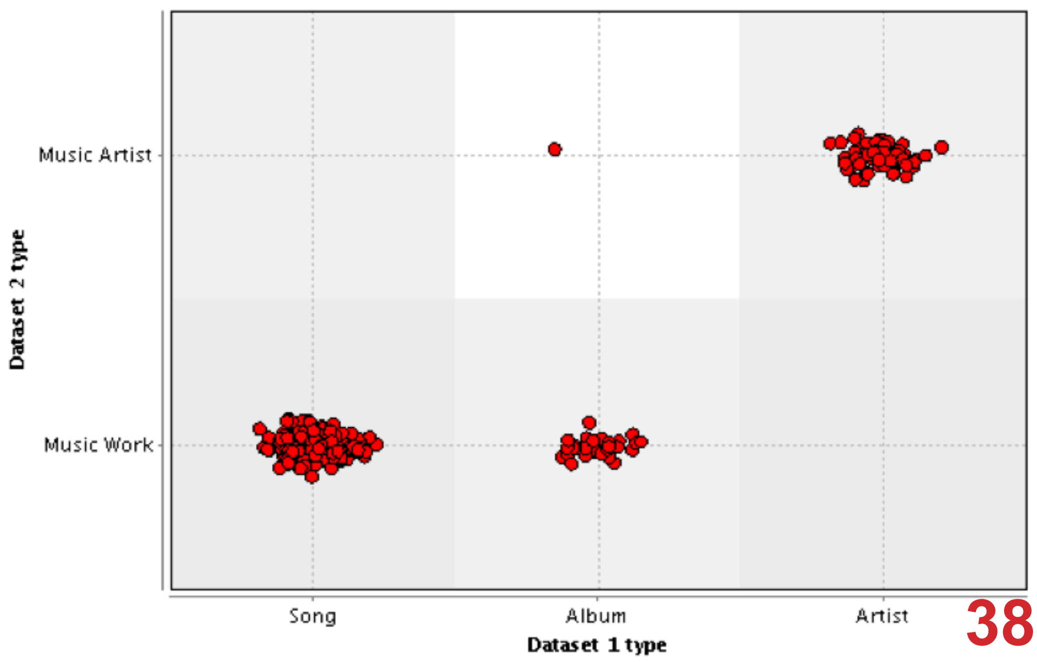


CONTENT BASED

[Paulheim, 2014]

Principle: links follow certain patterns, links that violate those patterns are erroneous.

- A multi-dimensional and scalable **outlier detection** approach for finding **erroneous identity links**.
- Projection of links into **Vector Space**: each link is a point in an n-dimensional vector space

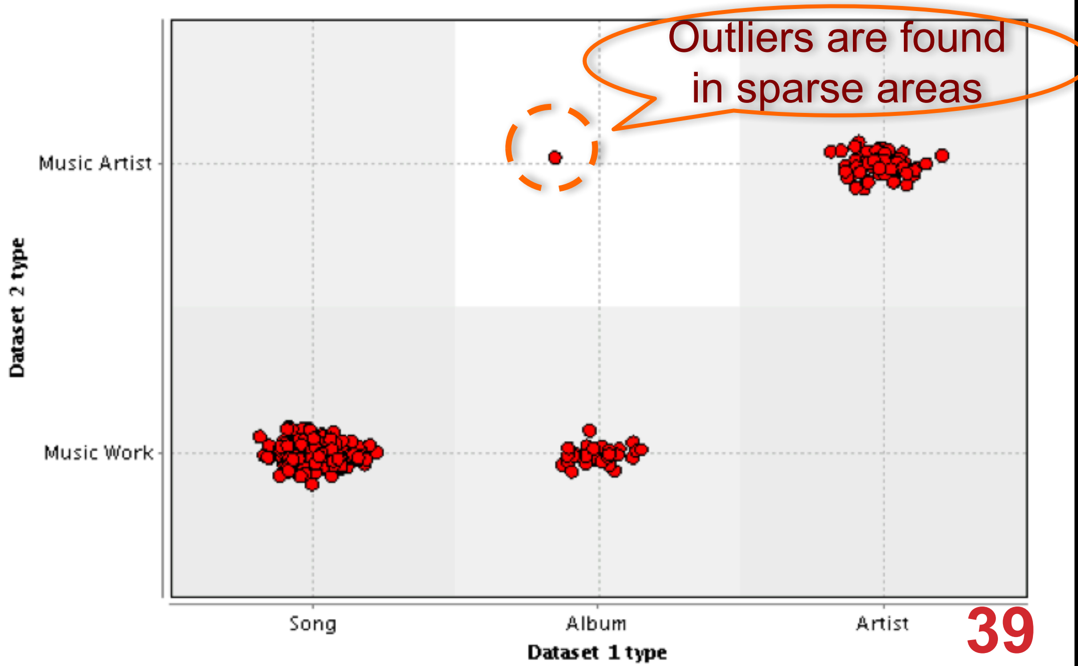


CONTENT BASED

[Paulheim, 2014]

Principle: links follow certain patterns, links that violate those patterns are erroneous.

- A multi-dimensional and scalable **outlier detection** approach for finding **erroneous identity links**.
- Projection of links into **Vector Space**: each link is a point in an n-dimensional vector space



CONTENT BASED

[Paulheim, 2014]

- **Feature Vector**: resource types and ingoing/outgoing properties
  - e.g. LHS_foaf:based_near and RHS_foaf:based_near are distinct features.

- **Different strategies of creating vectors**: direct types only, all ingoing and outgoing properties, or a combination
- Several outlier detection methods were tested: LOF, CBLOF, LOP, 1-class SVM etc.
- Each method assign a score to each data point indicating the likeliness of being an outlier ➔ **incorrect link**.

CONTENT BASED

[Paulheim, 2014]

- **Dataset**

	D1		D2	
Dataset	Peel Session	DBpedia	DBTropes	DBpedia
# Links	2,087		4,229	
# Types	3	31	2	79
# Properties	4	56	18	124

- **Gold Standard**: 100 randomly sampled links from D1 and D2
- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)

CONTENT BASED

[Paulheim, 2014]

- **Dataset**

Dataset	D1 Peel Session	D1 DBpedia	D2 DBTropes	D2 DBpedia
# Links	2,087		4,229	
# Types	3	31	2	79
# Properties	4	56	18	124

- **Gold Standard**: 100 randomly sampled links from D1 and D2
- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)
- **Best performance on D1:**
  - CBLOF (F1= **0.537**),1-class SVM (AUC = **0.857**)
- **Best performance on D2:**
  - LOF (F1= **0.5**, AUC = **0.619**)

CONTENT BASED

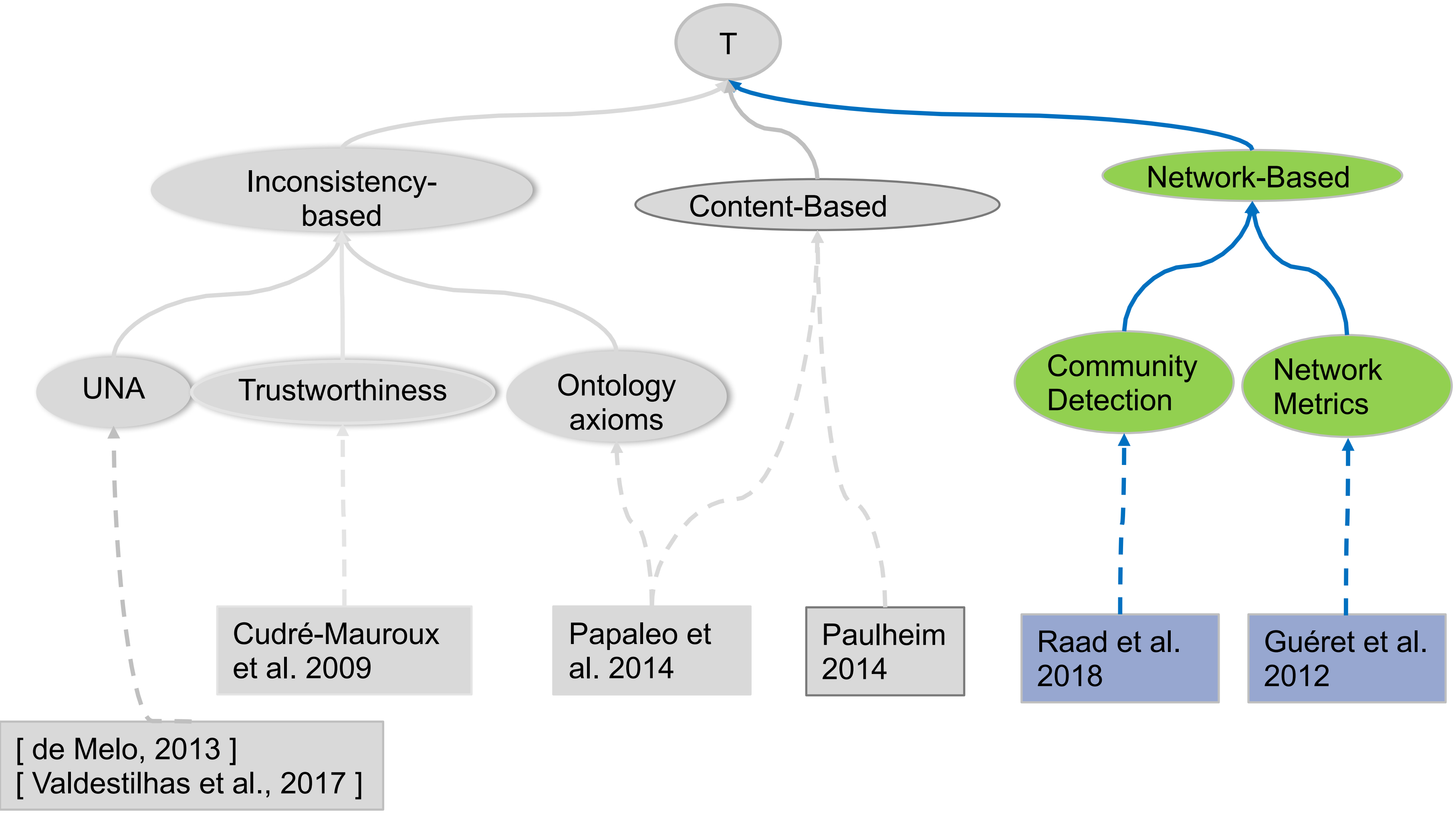
[Paulheim, 2014]

- **Dataset**

	D1		D2	
Dataset	Peel Session	DBpedia	DBTropes	DBpedia
# Links	2,087		4,229	
# Types	3	31	2	79
# Properties	4	56	18	124

- **Gold Standard**: 100 randomly sampled links from D1 and D2
- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)
- **Best performance on D1:**
  - CBLOF (F1= **0.537**),1-class SVM (AUC = **0.857**)
- **Best performance on D2:**
  - LOF (F1= **0.5**, AUC = **0.619**)
- Examples of **typical source of errors** for D1:
  - Linking of songs to albums with the same name.
  - Linking of different persons of the same name,
    e.g., a blues musician named Jimmy Carter to the U.S. president.

1. DETECTION OF ERRONEOUS IDENTITY LINKS



NETWORK BASED

[Guéret *et al.*, 2012]

[Raad *et al.*, 2018, UR]

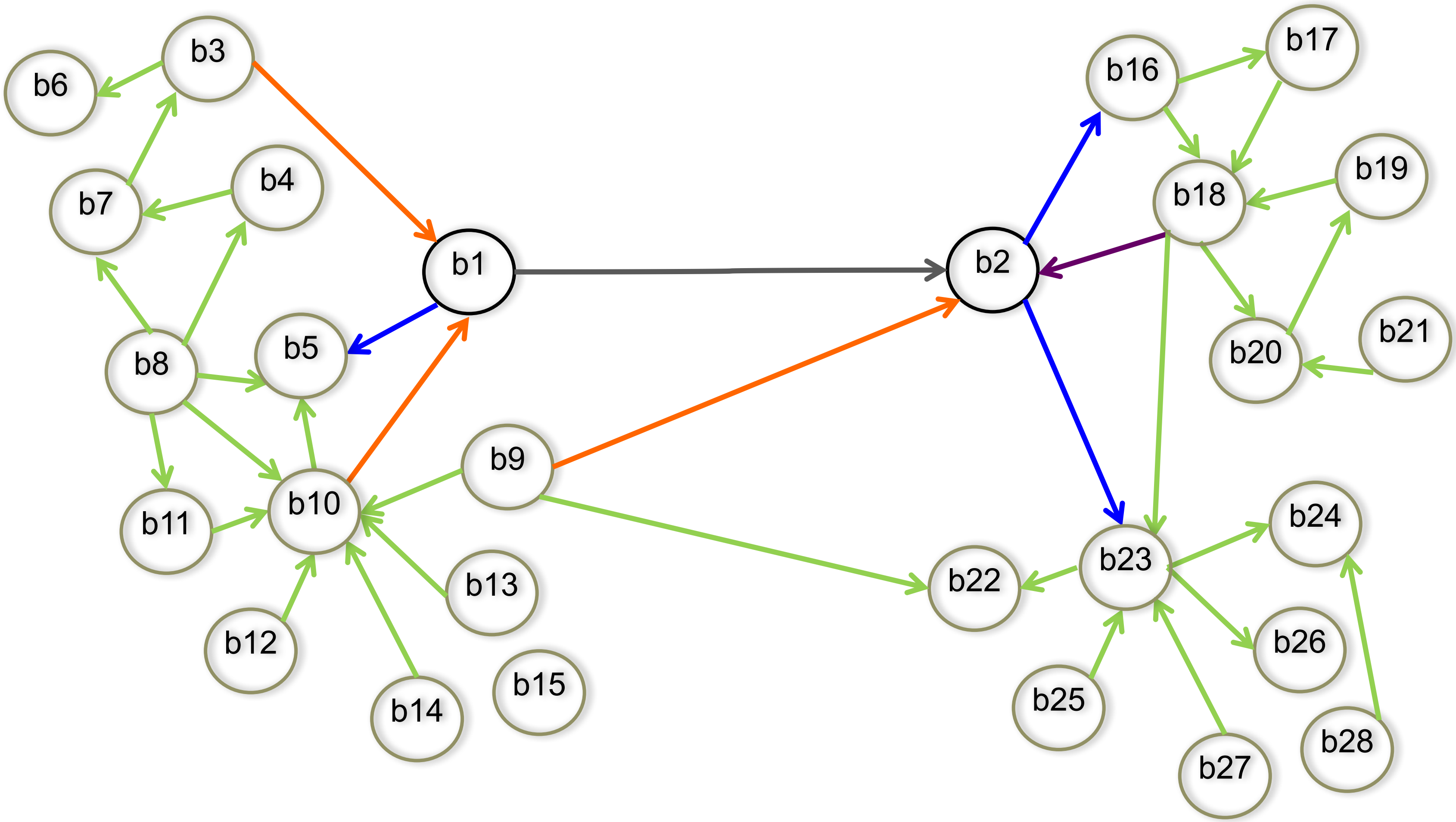
Principle

- The quality of a link can be determined based on **how connected a node** is within the **network** in which it appears.
- Use of **network metrics and structures** can help to detect erroneous links?

NETWORK BASED

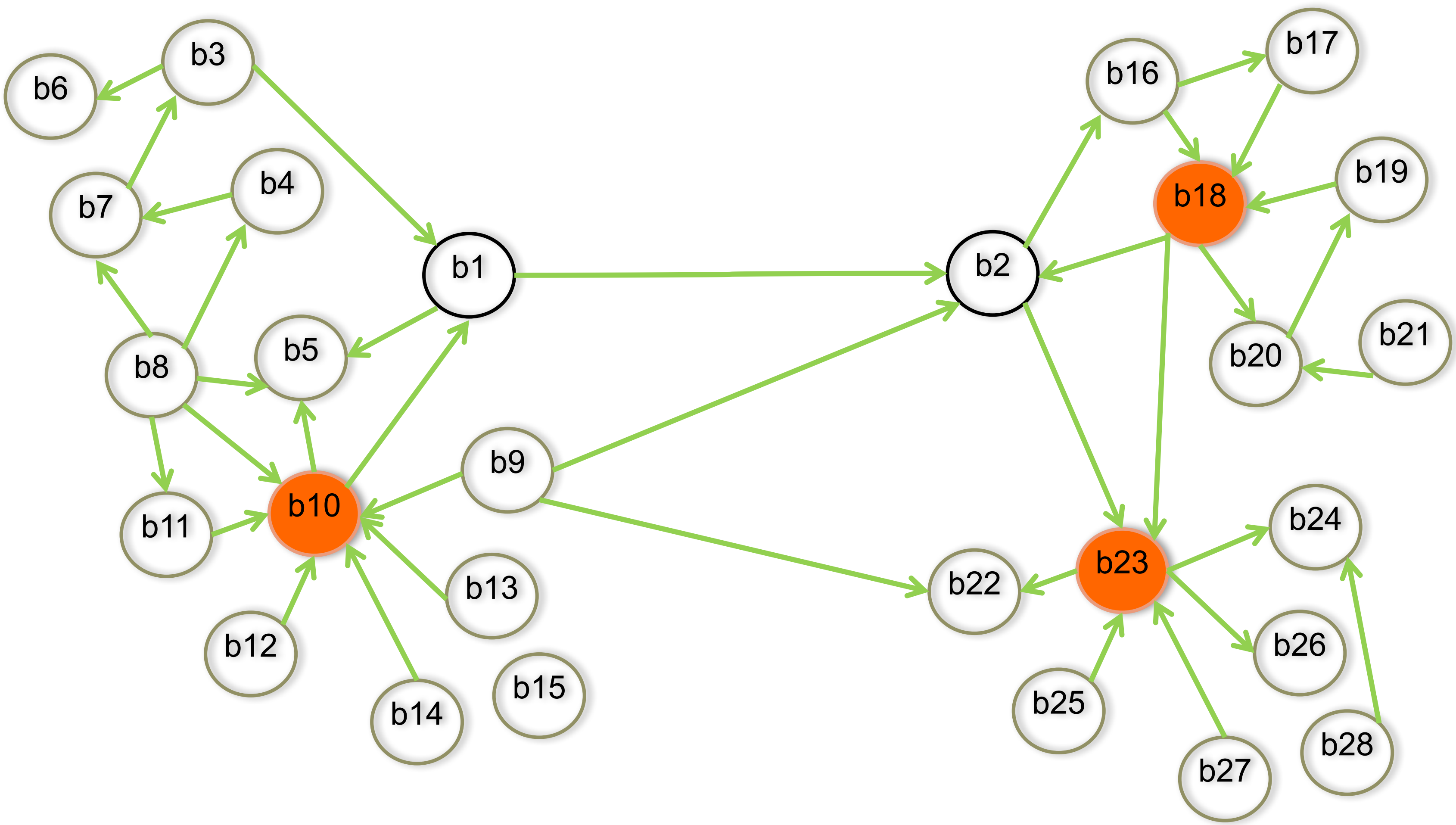
Node in-degree and out-degree

Clustering coefficient:
C(b16)=1; C(b20)=2/3; C(b23)=4/7



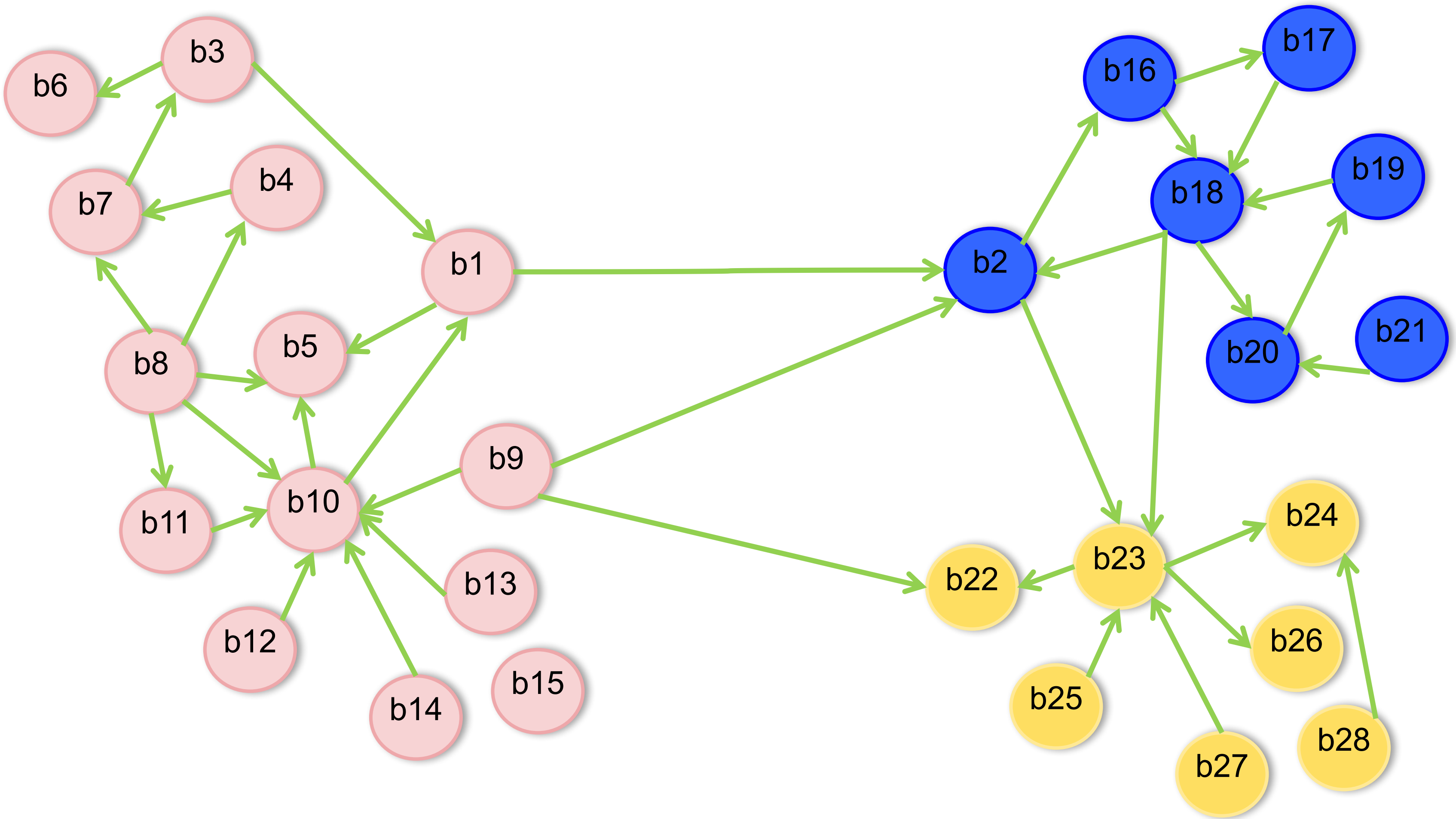
NETWORK BASED

Centrality



NETWORK BASED

Modularity and communities



NETWORK BASED

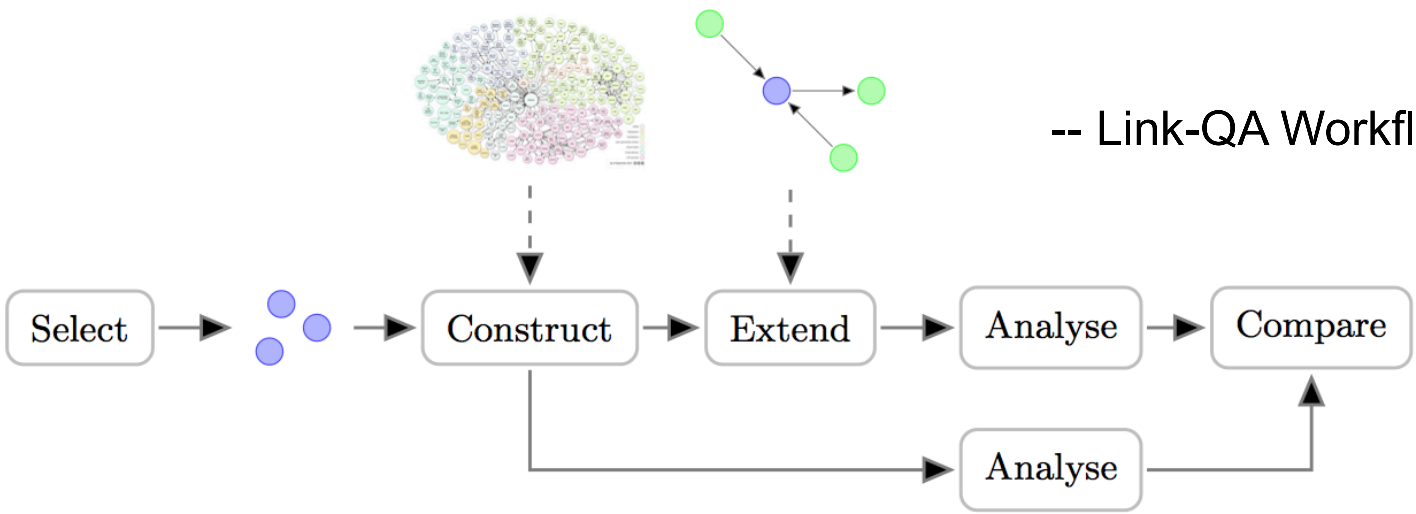
[Guéret *et al.*, 2012]

- Use of **network metrics** can help to detect erroneous links?
- Changes in quality of the Web of Data with the introduction of new links between datasets.

- **It is based on the use of**
  - three classic network metrics: clustering coefficient, centrality and degree
  - two Linked Data-specific ones: owl:sameAs chains, and description richness

NETWORK BASED

[Guéret *et al.*, 2012]



-- Link-QA Workflow

- The approach selects a set of resources and constructs a local network for each resource by querying the Web of Data.
- After analysis, i.e., measuring the different metrics, each local network is extended by adding new edges and analyzed again.
- The result coming from both analyses are compared to ideal distribution for the different metrics.

NETWORK BASED

[Guéret *et al.*, 2012]

Dataset

- The European project LOD Around the Clock (LATC) aims to enable the use of the Linked Open Data cloud for research and business purposes.

- LATC created a set of linking specifications (link specs) for Silk engine
  - 6 link sets are selected containing more than 50 correct and incorrect links

    e.g., geonames-linkedGeodataMountain, linkedct-pubmedDisease, …

- **Samples** taken from the **generated links** are manually **checked**

  ➔ Two reference sets containing all the positive (correct, good) and negative (incorrect, bad) links of the sample.

NETWORK BASED

[Guéret *et al.*, 2012]

Evaluation questions

- Do **positive** linksets **decrease** the distance to a metric's defined **ideal**, whereas **negative** ones **increase** it?
  - If that is the case, it would allow us to distinguish between link sets having **high** and **low ratios of bad links**.
- Is there a correlation between **outliers** and **bad links**?
  - If so, resources that rank **farthest** from the **ideal** distribution of a metric would relate to **incorrect links** from/to them.

NETWORK BASED

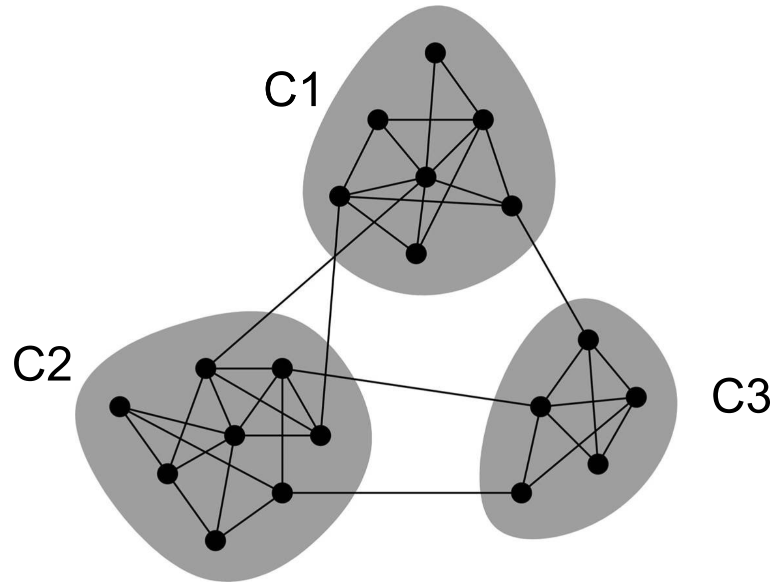
[Guéret *et al.*, 2012]

- **Recall = 0.68**
- **Precision = 0.49**

- **Conclusion:**
  - Common metrics such as centrality, clustering, and degree are **insufficient** for **detecting quality.**
  - **Description Richness** and **Open SameAs Chain** metrics look more promising, especially at detecting good and bad links, but they report too **many false positives.**

NETWORK BASED

[Raad *et al.*, 2018, under review]



- Considers the **identity network** build from the **explicit identity network** of sameAs links: removing of symmetric and reflexive links.
- Uses of Louvain **community detection** algorithm to detect subgraphs in the **identity network** that are highly connected.
- Defines a **ranking score** for each (intra-community and inter-community) identity link based on the **density of the community**.

NETWORK BASED

[Raad *et al.*, 2018, under review]

Ranking of identity links

intra-community erroneousness degree

$$a)\ err(e_C) = \frac{1}{w(e_C)} \times \left(1 - \frac{W_C}{|C| \times (|C| - 1)}\right)$$

inter-community erroneousness degree

$$b)\ err(e_{C_{ij}}) = \frac{1}{w(e_{C_{ij}})} \times \left(1 - \frac{W_{C_{ij}}}{2 \times |C_i| \times |C_j|}\right)$$

Correct link 0 — Error degree — 1 Erroneous link

NETWORK BASED

[Raad *et al.*, 2018, under review]

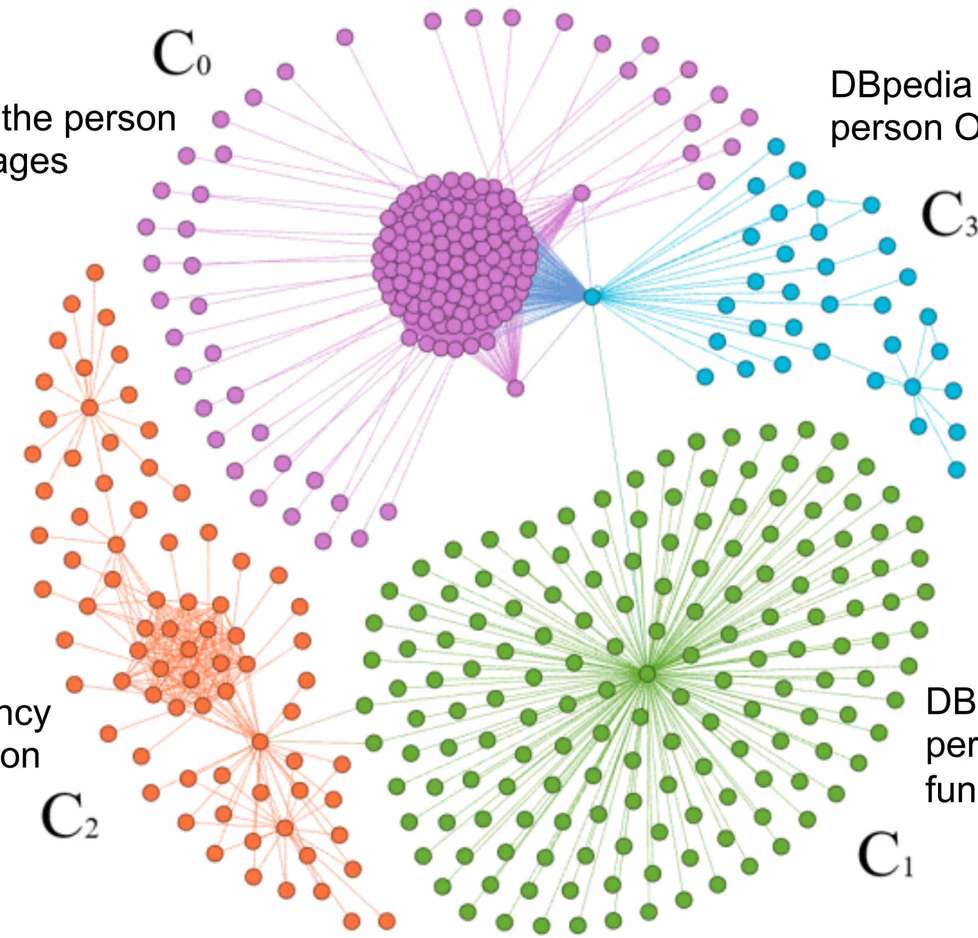
Dataset

- LOD-a-lot dataset [Fernandez et al. 2017]: a compressed data file of 28B triples from LOD 2015 crawl
- An **explicit identity network** of 558.9M edges (links) and 179M nodes (resources)
- Identity network of **331M** edges and **179M** nodes: after removing symmetric and reflexive links.

NETWORK BASED

[Raad *et al.*, 2018, under review]

Barack Obama's Equality Set

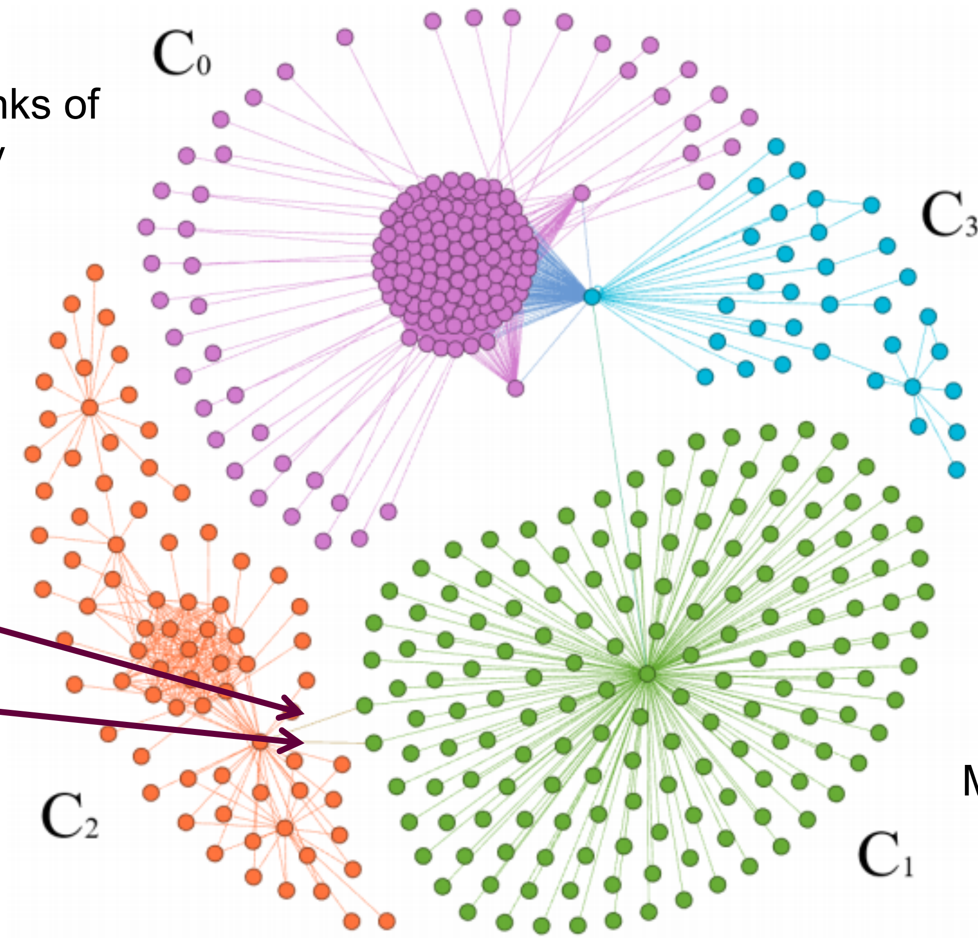


NETWORK BASED

[Raad *et al.*, 2018, under review]

Barack Obama's Equality Set

Low err(e) for the links of this community



These two links have err(e)= 1

Most of the links have err (e) = 0.9

NETWORK BASED

[Raad *et al.*, 2018, under review]

Precision on a randomly chosen set identity links from LOD

	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	total
same	35(100%)	22(100%)	18(85.7%)	7(77.7%)	15(68.1%)	**97**(88.9%)
related	0	0	2	2	2	**6**
unrelated	0	0	1	0	5	**6**
related + unrelated	0(0%)	0(0%)	3(14.2%)	2(22.2%)	7(31.8%)	**12**(11%)
can't tell	5	18	19	31	18	**91**
Total	**40**	**40**	**40**	**40**	**40**	**200**

- **Scales up** to a graph of **28.3 billion** triples: **12 hours**
- **Validates correct owl:sameAs links**
  - 100% of owl:sameAs with an **erroneousness degree <0.4** are correct
- Can **invalidate a large set of owl: sameAs links** on the LOD:
  - **1.26M** owl:sameAs have an **erroneousness degree** in [0.99, 1]

ERRONEOUS LINK DETECTION: SUMMARY

Positive points

- Different approaches relaying on different kinds of information (constraints, axioms, content and network)
- Good scalability of the approaches: up to 28.3 Billion triples
- Evaluations on real data on the LOD

ERRONEOUS LINK DETECTION: SUMMARY

Positive points

- Different approaches relaying on different kinds of information (constraints, axioms, content and network)
- Good scalability of the approaches: up to 28.3 Billion triples
- Evaluations on real data on the LOD

Limitations

- **Qualitative evaluation** often missing or conducted on only insignificant number of links (**max= 200** over **331M**)
- Some **assumptions** can be assumed on only **few datasets** on the LOD: UNA and provenance information.
- **Ontology axioms** are not always **available**: how to ensure their **validity** in every dataset. Is the LocatedIn is functional for every museum?
- **Difference** relationships are rarely available: useful for inconsistency checking

ERRONEOUS LINK DETECTION: SUMMARY

*"common metrics such as centrality, clustering, and degree **are insufficient for detecting quality** … Description Richness and Open SameAs Chain metrics look more promising, especially at detecting good and bad links, respectively, they report too **many false positives** for reference sets"* [Gueret et al., 2012]	*"Data linking algorithms (LIMES, SILK and DBpedia Extraction Framework) have a better consistency index than repositories such as sameas.org (13%) "* [Valdestilhas et al., 2017]	*"Due to the subjectivity of near-identity and similarity, we suggest that additional properties be used to describe the exact nature of the relationship"* [de Melo 2013]
Need for hybrid approaches	Need for more controlled link publication protocols	Need for alternate links

2. USE OF ALTERNATE LINKS

2. USE OF ALTERNATE LINKS

Use of weaker alternative links to express relatedness between resources/concepts.

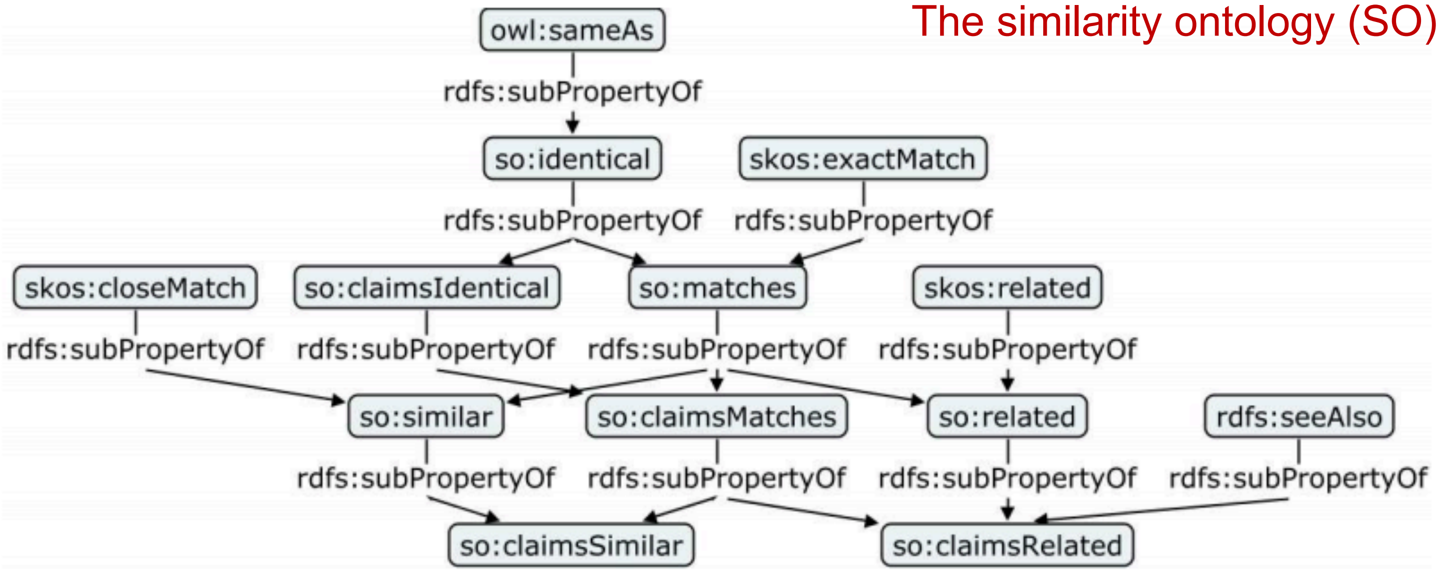
- UMBEL[1] vocabulary introduces umbel:isLike "*to assert a link between similar individuals who may be believed to be identical*"
- Vocab.org[2] introduces similarTo to be used when having two things that are not the owl:sameAs
- [de Melo, 2013] introduces lvont:nearlySameAs and lvont:somewhatSameAs, two predicates for expressing near-identity in the Lexvo.org[3]
- **Use of domain-specific identity relations:**
  - ex:sameBook to express identity between two books

[1] http://umbel.org
[2] http://vocab.org
[3] http://lexvo.org

2. USE OF ALTERNATE LINKS

- [Halpin et al., 2010] proposed a similarity ontology (SO) in which they hierarchically represent 13 different predicates including 8 new ones.

The similarity ontology (SO)



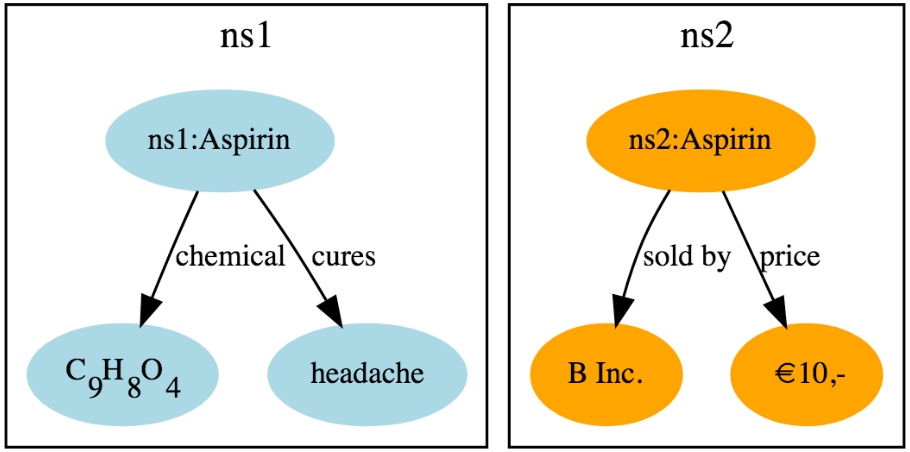
		Transitive	Non-transitive
Reflexive	Symmetric	*so:identical*	*so:similar*
	Non-Symmetric	*so:claimsIdentical*	*so:claimsSimilar*
Non-Reflexive	Symmetric	*so:matches*	*so:related*
	Non-Symmetric	*so:claimsMatches*	*so:claimsRelated*

Reflexivity, Symmetry and Transitivity properties for the 8 new predicates.

3. CONTEXTUAL IDENTITY LINKS

3. CONTEXTUAL IDENTITY LINKS

- **Weaker** kinds of **identity** can be expressed by considering a **subset of properties** with respect to which two resources can be considered to be the same.
- Identity is **context-dependent** [Geach, 1967]
  - *allowing two medicines to be considered the same in terms of their chemical substance, but different in terms of their price (e.g., because they* are produced by different companies).



3. CONTEXTUAL IDENTITY LINKS

[Beek et al., 2016]

- Propose an approach that allows the **characterization o**f the **context** in which an **identity link is valid**

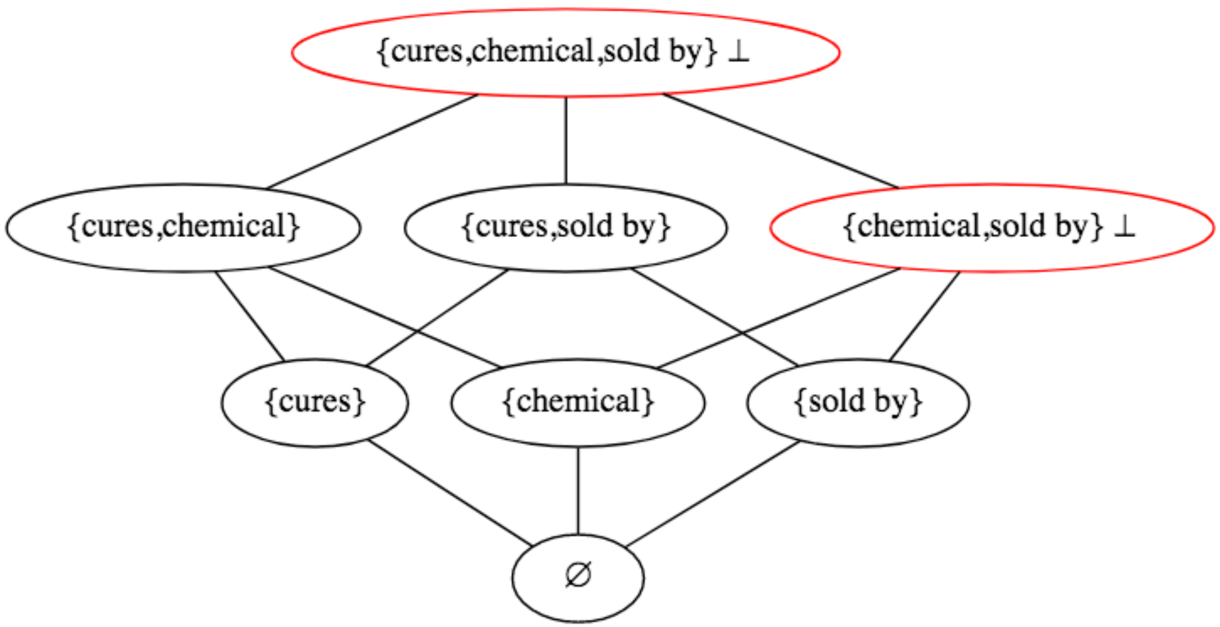
- A context is a **subset of properties** for which two individuals must have the **same values**

- Contextual identity link **preserves equivalence relation**, w.r.t. a subset of the properties

3. CONTEXTUAL IDENTITY LINKS

[Beek et al., 2016]

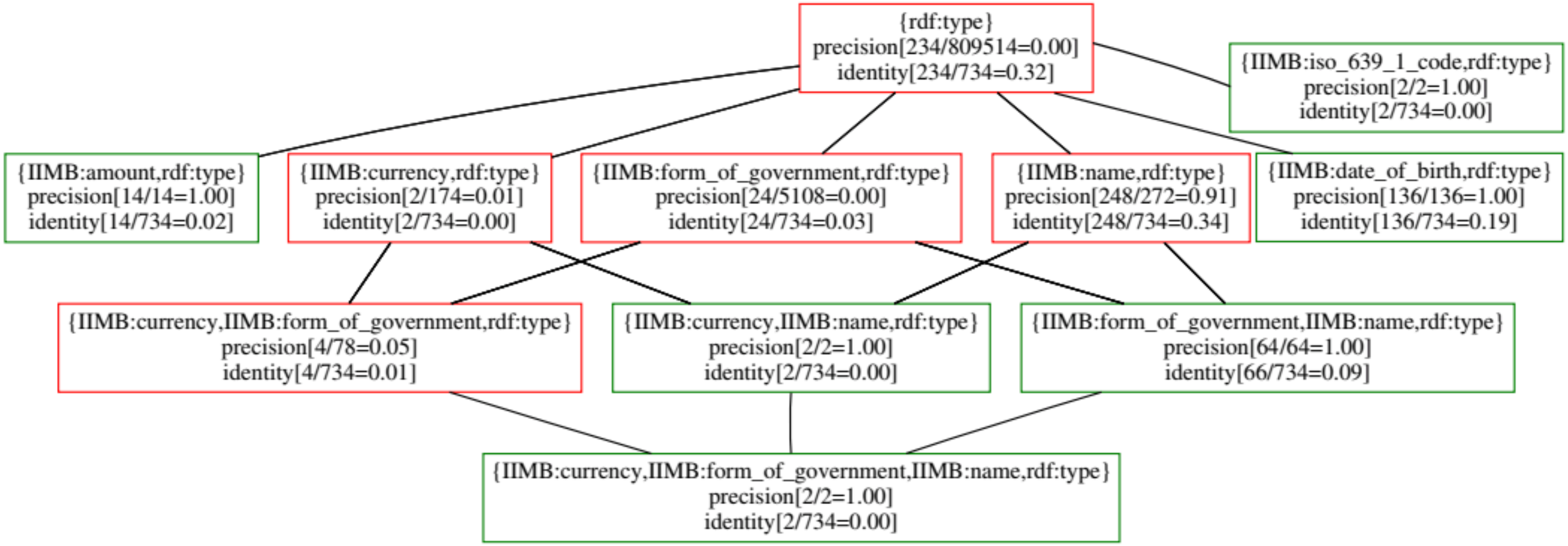
- All the **possible subsets** of properties organized in a **lattice** using the set inclusion relation.



3. CONTEXTUAL IDENTITY LINKS

[Beek et al., 2016]

- Evaluation on a dataset in the instance matching track of the OAEI2012 : a variant of the IIMB datasets.
- The obtained identity subrelations



3. CONTEXTUAL IDENTITY LINKS

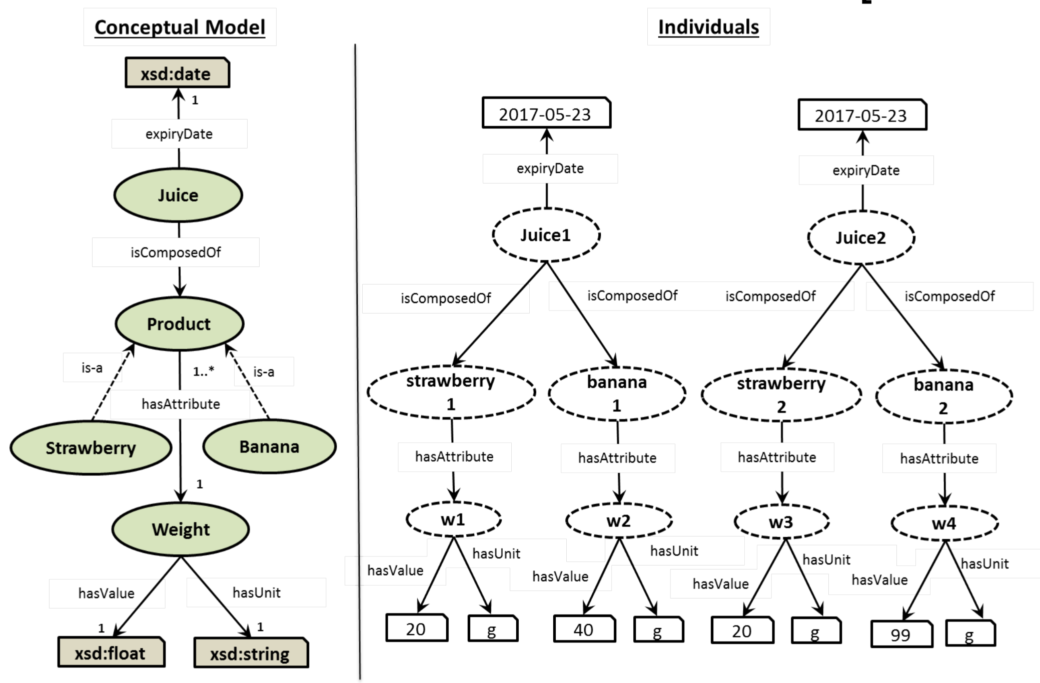
[Raad et al., 2017]

- New predicate *:identiConTo* for expressing **contextual identity** relation
- An **algorithm** for automatic detection of the **most specific contexts** in which two instances (resources) are identical
  - the detection process can further be guided by a set of **semantic constraints** that are provided by domain experts.

- Contexts are defined as a sub-ontology of the domain ontology
- All the possible contexts are organized in a lattice using an order relation.

3. CONTEXTUAL IDENTITY LINKS

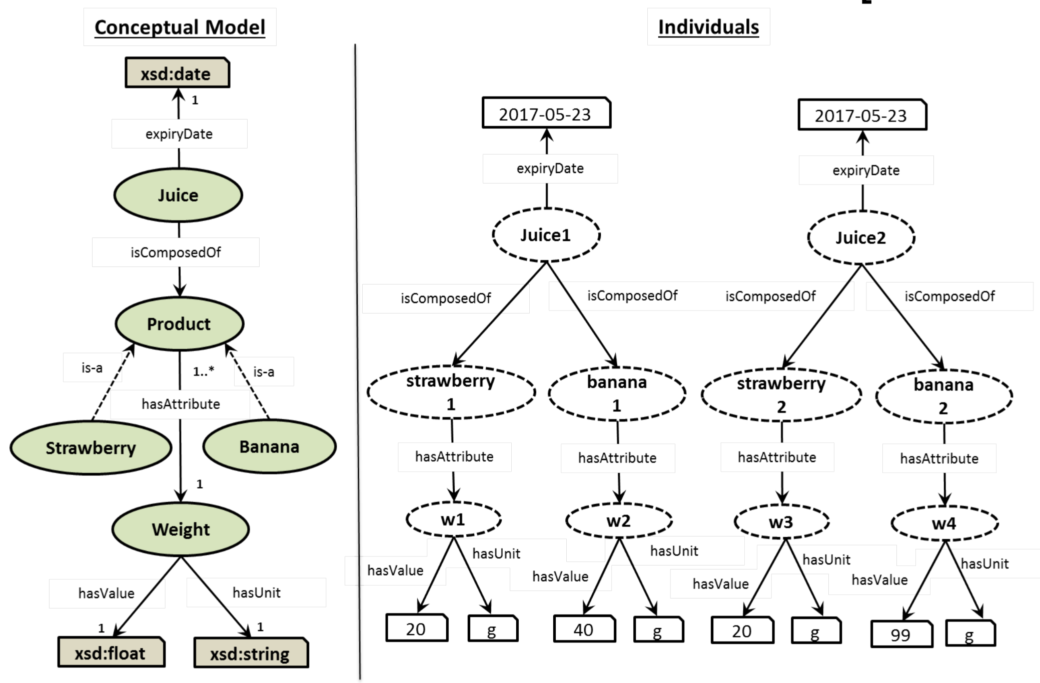
[Raad et al., 2017]



3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Contexts are defined as a **sub-ontology** of the domain ontology



Contextual Identity Link Example

Π_a(Juice) = { (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf $^{-1}$}),
(Strawberry, {rdf:Type}, {hasAttribute, isComposedOf $^{-1}$}),
(Weight, {rdf:Type, hasValue, hasUnit}, {hasAttribute^{-1}}) **}**

***identiConTo*$_{<\Pi_a(\text{Juice})>}$(juice1, juice2)**

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Π_a(Juice) = { (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf $^{-1}$}),
(Strawberry, {rdf:Type}, {hasAttribute, isComposedOf $^{-1}$}),
(Weight, {rdf:Type, hasValue, hasUnit}, {hasAttribute^{-1}}) **}**

Π_b(Juice) = { (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {hasAttribute, isComposedOf $^{-1}$}),
(Strawberry, {rdf:Type}, {hasAttribute, isComposedOf $^{-1}$}),
(Weight, {rdf:Type, hasUnit}, {hasAttribute^{-1}}) **}**

Π_c(Juice) = { (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf $^{-1}$}),
(Strawberry, {rdf:Type}, {isComposedOf $^{-1}$}) **}**

3. CONTEXTUAL IDENTITY LINKS

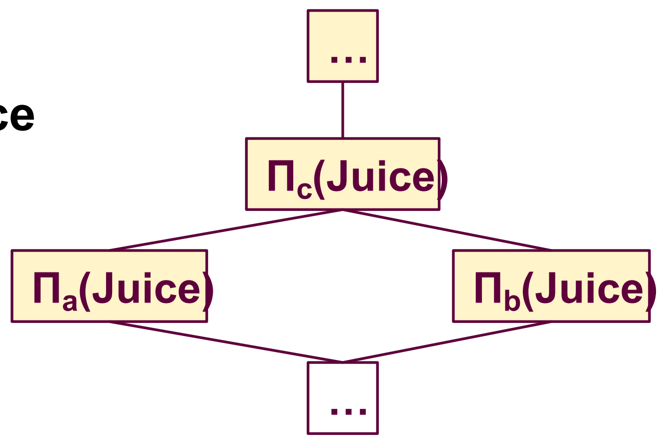
[Raad et al., 2017]

Π_a(Juice) = { (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf^{-1}}),
(Strawberry, {rdf:Type}, {hasAttribute, isComposedOf^{-1}}),
(Weight, {rdf:Type, hasValue, hasUnit}, {hasAttribute^{-1}}) **}**

Π_b(Juice) = { (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {hasAttribute, isComposedOf^{-1}}),
(Strawberry, {rdf:Type}, {hasAttribute, isComposedOf^{-1}}),
(Weight, {rdf:Type, hasUnit}, {hasAttribute^{-1}}) **}**

Π_c(Juice) = { (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf^{-1}}),
(Strawberry, {rdf:Type}, {isComposedOf^{-1}}) **}**

The **possible contexts** are organized in a **lattice** using an order relation.



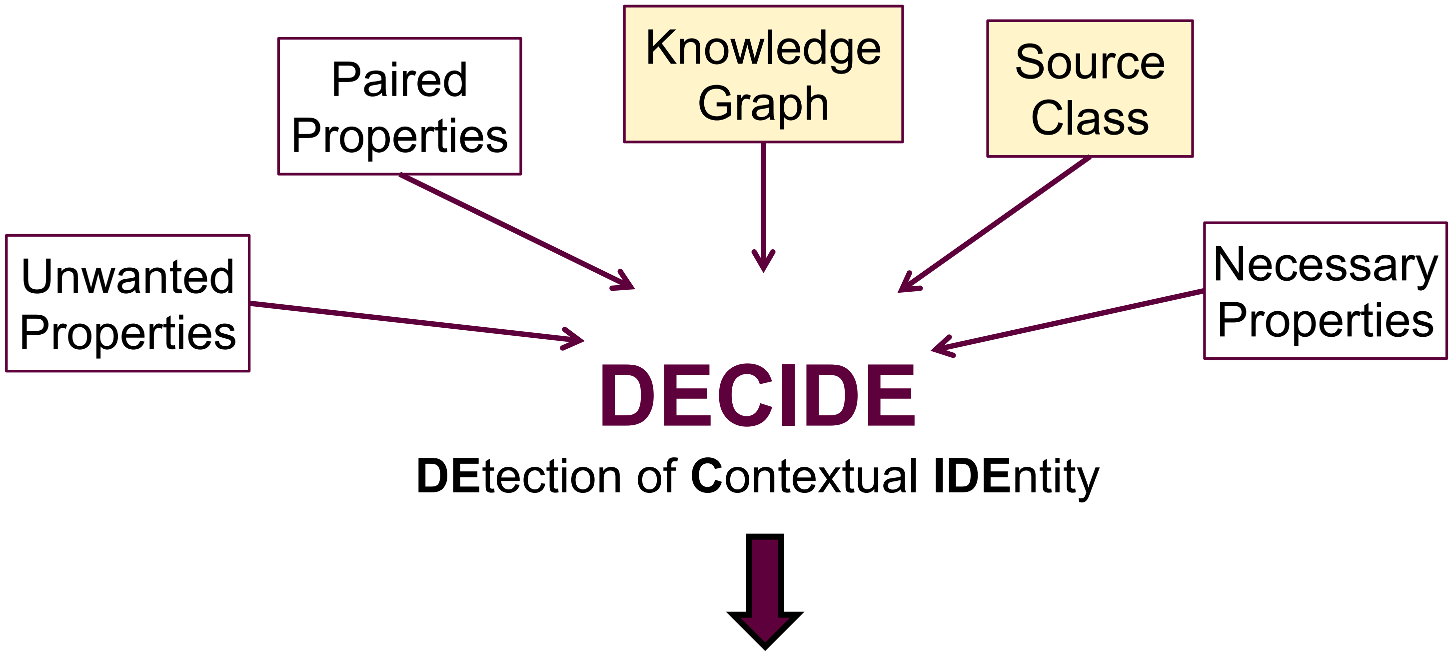
Π_a(Juice) ≤ Π_c(Juice) **Π_b(Juice) ≤ Π_c(Juice)**

each local context in Π_c(Juice) is less specific or equal to its corresponding local context in Π_a(Juice)

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

It automatically detects and adds these contextual identity links in the knowledge graph



For each pair of instances (i_1, i_2) of the source class
set of the most specific global contexts in which (i_1, i_2) are identical

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Transformation of Micro-organisms

Digestion Process

PO²

A Process and Observation Ontology

RDF

PO²

A Process and Observation Ontology

CellExtraDry

- **Classes:** ≈ 4 700
- **Individuals:** ≈ 415 000
- **Statements:** ≈ 1 700 000

PO²

A Process and Observation Ontology

Carredas

RDF

- **Classes :** ≈ 5 000
- **Individuals:** ≈ 42 000
- **Statements :** ≈ 237 000

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

PO²
A Process and Observation Ontology

CellExtraDry + Carredas

- 950 *classes*
- 1,5 million *triplets*
- 284 *processus de transformations*

Irrelevant properties
up = (observation, observes, *)

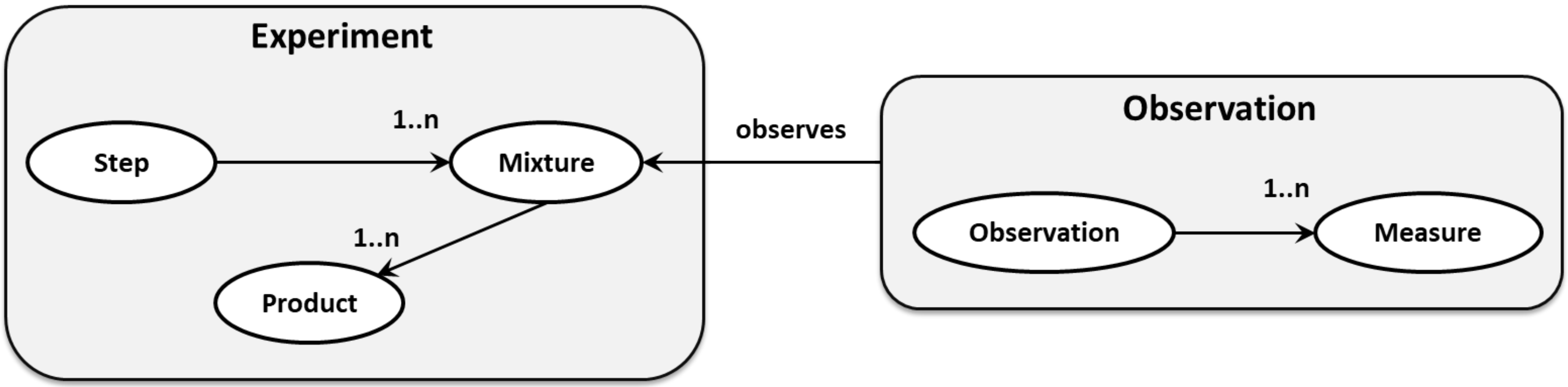
DECIDE

DEtection of **C**ontextual **IDE**ntity

	Experiment 1	Experiment 2
	Mixture	Step
# Instances	1,187	581
# Possible pairs	703,891	168,490
# Distinct Global Contexts	2 232	718
# Contextual identity links	1, 279,376	348,017
# Contextual identity links per pair	1.81	2.06

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]



Detect for each context $\mathbf{GC_i}$, the measures $\mathbf{m_i}$ where

$$\textbf{\textit{identiConTo}}_{<GCi>}(i_1, i_2) \cap \textbf{\textit{observes}}(i_1, m_1) \rightarrow \textbf{\textit{observes}}(i_2, m_2)$$

with $m_1 \simeq m_2$

$$\textbf{\textit{identiConTo}}_{<GCi>}(\mathbf{i1, i2}) \rightarrow \mathbf{same(m_i)}$$

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Detection of 38 844 rules

Règle	*Taux d'erreur*	*Support*
$identiConTo_{<GC_1>}(x, y) \rightarrow$ same(pH)	6.19 %	57
$identiConTo_{<GC_3>}(x, y) \rightarrow$ same(Dureté)	1.86 %	66
$identiConTo_{<GC_2>}(x, y) \rightarrow$ same(Friabilité)	4.52 %	647

The domain experts has evaluated the plausibility of the best **20 rules** (in termes of error rate and support)

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Detection of 38 844 rules

Règle	*Taux d'erreur*	*Support*
$identiConTo_{<GC_1>}(x, y) \rightarrow \text{same(pH)}$	6.19 %	57
$identiConTo_{<GC_3>}(x, y) \rightarrow \text{same(Dureté)}$	1.86 %	66
$identiConTo_{<GC_2>}(x, y) \rightarrow \text{same(Friabilité)}$	4.52 %	647

The domain experts has evaluated the plausibility of the best **20 rules** (in termes of error rate and support)

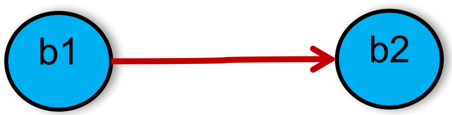
3	**5**	**4**	**5**	**3**	**plausibility**
Impossible	Not very probable	Can't tell	Why not	Plausible	

The error rate decreases of 12% when a global context is replaced by a more specific global context

LINK VALIDATION: SUMMARY

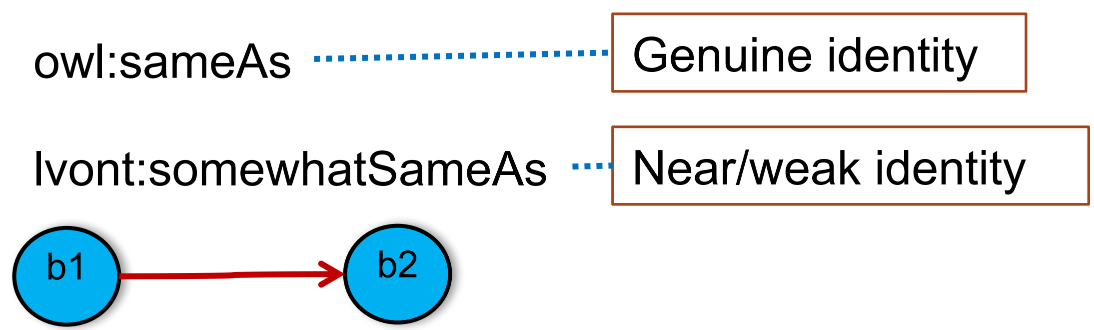
- Different kinds of identity relationship

....



LINK VALIDATION: SUMMARY

- Different kinds of identity relationship

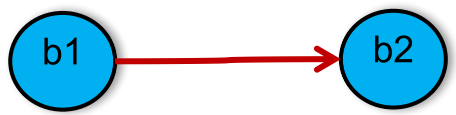


LINK VALIDATION: SUMMARY

- Different kinds of identity relationship

owl:sameAs — Genuine identity

lvont:somewhatSameAs — Near/weak identity



:sameBook

:sameExpression — Subjective identity

:sameArtWork

:dentiConTo$_{<\Pi a>}$ — Contextual identity

LINK VALIDATION: SUMMARY

- Different kinds of identity relationship
- Need of hybrid methods

Network Topology

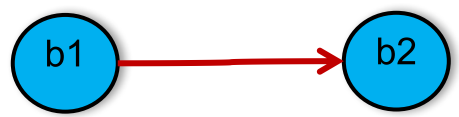
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

:dentiConTo$_{<\Pi a>}$

LINK VALIDATION: SUMMARY

- Different kinds of identity relationship
- Need of hybrid methods
- Link quality assessment is not a matter of one unique dimension

Network Topology

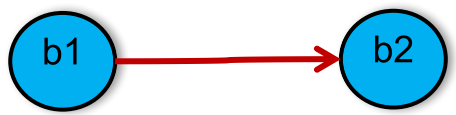
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

:dentiConTo$_{<\Pi a>}$

Link Validity: Inconsistent equivalent classes, Invalid links, Contextual links

Link Properties: Transitivity, symmetry, …

Link added-value: Information gain, reachability, …

Link meta-data: availability, evolution

LINK VALIDATION: SUMMARY

What is about the **distinctness** relation?

- Different kinds of identity relationship
- Need of hybrid methods
- Link quality assessment is not a matter of one unique dimension

Network Topology

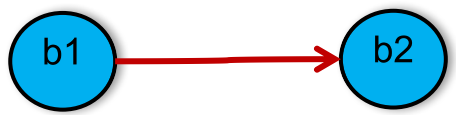
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

:dentiConTo$_{<\Pi a>}$

Link Validity: Inconsistent equivalent classes, Invalid links, Contextual links

Link Properties: Transitivity, symmetry, …

Link added-value: Information gain, reachability, …

Link meta-data: availability, evolution

REFERENCES (1)

[Beek et al., 2016] A contextualised semantics for owl: sameas.

W. Beek, S. Schlobach, and F. van Harmelen. In ESWC 2016

[CudreMauroux et al., 2009] idmesh: graph-based disambiguation of linked data.

P. CudreMauroux, P. Haghani, M. Jost, K. Aberer, and H. De Meer. In WWW 2009.

[de Melo, 2013] Not quite the same: Identity constraints for the web of linked data.

G. de Melo. In AAAI 2013.

[Geach, 1967] Identity. P. Geach. Review of Metaphysics, 21:3–12, 1967.

[Guéret et al. 2012] Assessing linked data mappings using network measures.

C. Guéret, P. Groth, C. Stadler, and J. Lehmann. In ESWC 2012

[Halpin et al., 2010] When owl:sameAs isn't the same: An analysis of identity in Linked Data.

H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. In ISWC 2010.

[Hogan et al., 2012] Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora.

A. Hogan, A. Zimmermann, J. Umbrich, A. Polleres, and S. Decker. In JWS 2012.

REFERENCES (2)

[Jaffri et al., 2008] URI disambiguation in the context of linked data.

A. Jaffri, H. Glaser, and I. Millard. In LDOW@WWW 2008.

[Paulheim, 2014] Identifying wrong links between datasets by multi-dimensional outlier detection.

H. Paulheim. In WoDOOM 2014.

[Papaleo et al., 2014] Logical detection of invalid sameas statements in rdf data.

L. Papaleo, N. Pernelle, F. Saïs, and C. Dumont. In EKAW 2014.

[Raad et al., 2017] Detection of contextual identity links in a knowledge base.

J. Raad, N. Pernelle, and F. Saïs. In K-CAP 2017.

[Raad et al., 2018 under review] Detecting Erroneous Identity Links on the Web using Network Metrics. J. Raad, W. Beek, F. van Harmelen, N. Pernelle and F. Saïs. Submitted to ISWC 2018

[Valdestilhas et al., 2017] Cedal: time-efficient detection of erroneous links in large-scale link repositories. A. Valdestilhas, T. Soru, and A.-C. N. Ngomo. In WI 2017.