TOWARDS A SOLUTION TO THE "SAMEAS PROBLEM"

Joe Raad

joe.raad@agroparistech.fr

July 12th, 2018 - DIG Seminar

ABOUT ME

PHD STUDENT

* 3rd year
* MIA-Paris (INRA, AgroParisTech)
* LRI (CNRS)

Interest: Managing Identity in the Semantic Web

Website: www.joe-raad.com

MOTIVATION

5 ★ LINKED OPEN DATA

★ make your data available on the Web

★★ make it available as structured data

★★★ make it available in a non-proprietary format

★★★★ use open standards from the W3C

★★★★★ link your data to other data

Tim Berners-Lee, 2010

WHY LINKING YOUR DATA?

```
spotify:elvisPresley spotify:artistOf spotify:suspiciousMinds.
spotify:suspiciousMinds spotify:releaseDate "1969-01-01"^^xsd:
```

```
apple:artist_8723
        apple:birthday "1935-01-08"^^xsd:date;
        apple:bornIn usdata:tupelo-Mississipi.
```

Siri, play an American song from the late 60s

HOW TO LINK YOUR DATA?

owl:sameAs

(the semantic web identity predicate)

⟨x, owl:sameAs, y⟩

means that:

$$x = y$$

$$(\forall P)(Px \leftrightarrow Py)$$

there is one thing which has two names: x and y

WHY IDENTITY LINKS?

SIMILARITY IS NOT GOOD ENOUGH

> *“SKOS exactMatch indicates a high degree of confidence that two concepts can be used interchangeably across a wide range of information retrieval applications”*
> *SKOS specification, 2009*

<u>**NO FORMAL MEANING**</u>

CAN ONE ACTUALLY INFER ANYTHING FROM SAMEAS LINKS ON THE LOD?

(SPOILER: NOT SO MUCH)

1. **Difficulty in finding identical terms:** Like the WWW, the SW does not allow backlinks to be followed.

2. **Erroneous Inferences:** Like the WWW, the SW contains a great number of incorrect statements.

HOW TO FIX THIS?

1. Identity Service for the LOD to access:
    - the existing owl:sameAs statements
    - the list of identical terms

2. Detect the incorrect owl:sameAs links in the LOD

(Outline of this talk)

SAMEAS.CC

Identity Management Service in the LOD

SAMEAS.CC REQUIREMENTS

This solution must scale to the LOD Cloud.

This solution must be formally interpretable (no `skos:exactMatch`, `rdfs:seeAlso`).

It must be calculated incrementally.

FORMAL PROPERTIES OF IDENTITY

Identity is the smallest equivalence relation, it is:

- reflexive (x,x)
- symmetric (x,y) → (y,x)
- transitive (x,y) ∧ (y,z) → (x,z)

EXAMPLE

Explicit identity relation over `{:a, :b, :c, :d}`:

```
:a owl:sameAs :b
```

```
:d owl:sameAs :b
```

The closure results in two identity sets:

```
:a :b :d
```

```
:c
```

Then the implicit identity relation is:

```
:a owl:sameAs :a
:a owl:sameAs :b
:a owl:sameAs :d
:b owl:sameAs :a
:b owl:sameAs :b
```

```
:b owl:sameAs :d
:c owl:sameAs :c
:d owl:sameAs :a
:d owl:sameAs :b
:d owl:sameAs :d
```

APPROACH

3 MAIN STEPS

1. EXTRACT THE EXPLICIT IDENTITY STATEMENTS

INPUT: LOD-a-lot = 28.3B triples (Fernandez et al., 2017)

```
prefix owl: <http://www.w3.org/2002/07/owl#>
        select distinct ?s ?p ?o {
          bind (owl:sameAs ?p)
          ?s ?p ?o
        }
```

OUTPUT: 558.9M owl:sameAs (179.7M terms)

2. COMPACT THE EXPLICIT IDENTITY STATEMENTS

INPUT: 558.9M owl:sameAs (179.73M terms)

GNU sort unique:

leaves out 2.8M reflexive triples

leaves out 225M duplicate symmetric triples

OUTPUT: 331M owl:sameAs (179.67M terms)

3. CALCULATE THE IMPLICIT IDENTITY RELATION

INPUT: 331M owl:sameAs (179.67M terms)

Assign each term to an identity set
(algorithm described in the paper)

OUTPUT: 48.9M non-singleton identity sets

SOME STATS

- This approach takes around 10 hours using 2 CPU cores on a regular SSD disk laptop
- 558.9M sameAs → 48.9M non-singleton identity sets
- 64% of identity sets have cardinality of 2
- Materialization consists of 35.2B sameAs triples

WHAT WE DID TILL NOW

- Provided the largest dataset of semantic identity links to date
- Presented an efficient approach for calculating and storing the closure of these links
- Provided a resource (http://sameas.cc) for querying and downloading the data
- Provided several analytics over the data and the usage of identity in the LOD (check our paper)

WHY WE DID IT?

- Findability of backlinks
- Query answering
- Query answering under entailment
- Verification of the correctness of the identity links

USE CASE

The largest identity set contains 177,794 terms

Meaning

there is 177,794 names (IRIs) that refers to the same real world entity

Reality

full list at: https://sameas.cc/term?id=4073

```
http://dbpedia.org/resource/Albert_Einstein
http://dbpedia.org/resource/Basketball
http://dbpedia.org/resource/Coca-Cola
http://dbpedia.org/resource/Deauville
http://dbpedia.org/resource/Italy
...
```

DETECTION OF ERRONEOUS IDENTITY LINKS

HOW CAN WE DETECT ERRONEOUS SAMEAS LINKS?

Source Trustworthiness

[Cudre-Mauroux et al. 2009]

UNA or Ontology Axioms Violation

[de Melo 2013; Valdestilhas et al. 2017; Hogan et al. 2012; Papaleo et al. 2014]

Content-based

[Paulheim et al. 2014 ; Cuzzola et al.,2015]

Network Metrics

[Guéret et al. 2012]

WHAT WE NEED

High accuracy and recall

Tested on real world data

Scalable to the LOD

Not require any assumption on the data
(e.g. UNA, textual description, source trustworthiness)

(No existing approach combines all these criteria)

APPROACH

Use the community structure of the network containing solely sameAs links to assign an error degree for each link

4 MAIN STEPS

1. EXTRACT THE EXPLICIT IDENTITY STATEMENTS

INPUT: LOD-a-lot = 28.3B triples
(Fernandez et al., 2017)

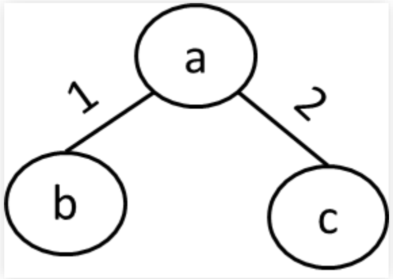
```
prefix owl: <http://www.w3.org/2002/07/owl#>
        select distinct ?s ?p ?o {
          bind (owl:sameAs ?p)
          ?s ?p ?o
        }
```

OUTPUT: 558.9M owl:sameAs (179.7M terms)

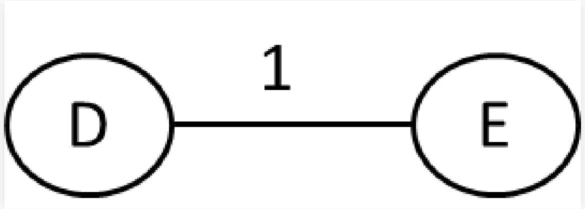
2. PARTITION TO EQUALITY SETS

```
:a owl:sameAs :b
:a owl:sameAs :c
:c owl:sameAs :a
:d owl:sameAs :e
```

Eq Set 1



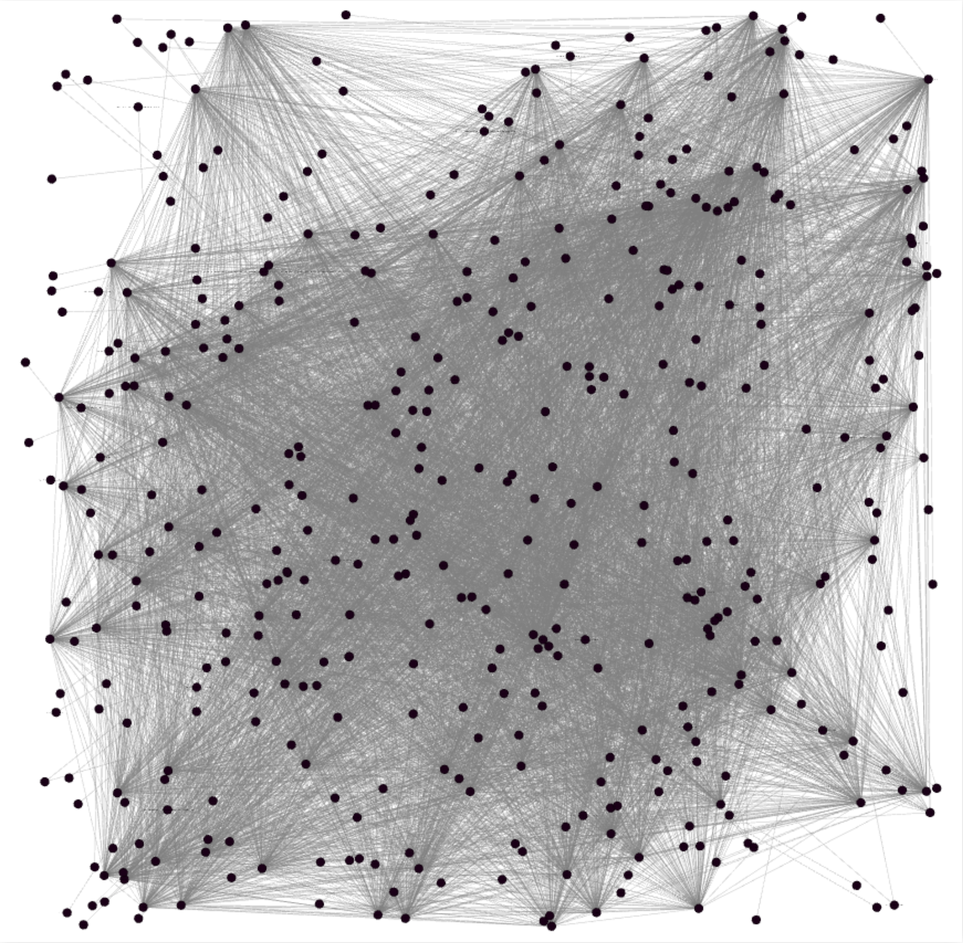
Eq Set 2



48.9M equality sets total

'BARACK OBAMA' EQUALITY SET

These identifiers denote the exact same thing (EqSet 5723)



3. DETECT THE COMMUNITY STRUCTURE IN EACH EQ SET

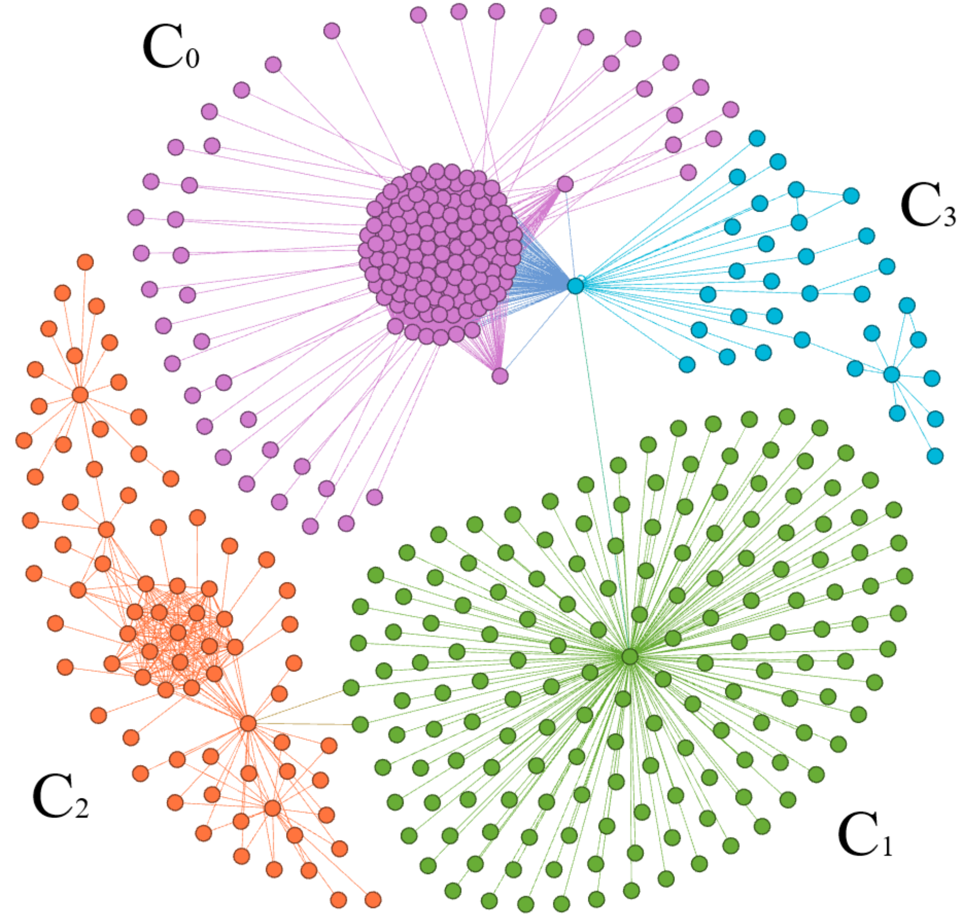
We use the Louvain algorithm [Blondel et al. 2008]

- Detects non-overlapping communities
- Adapted to weighted networks
- Linear computational complexity
- Outperforms other algorithms

  [Lancichinetti and Fortunato. 2009 ; Yang et al. 2016]

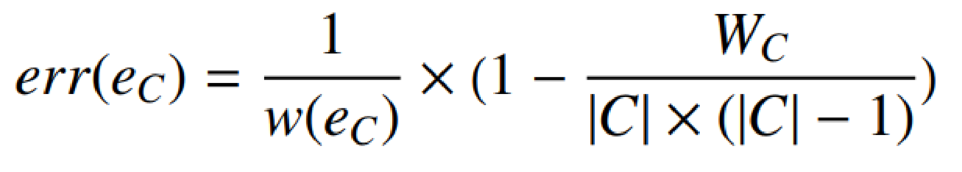
COMMUNITIES - 'BARACK OBAMA'

C0: person; C1: president; C2: government; C3: senator



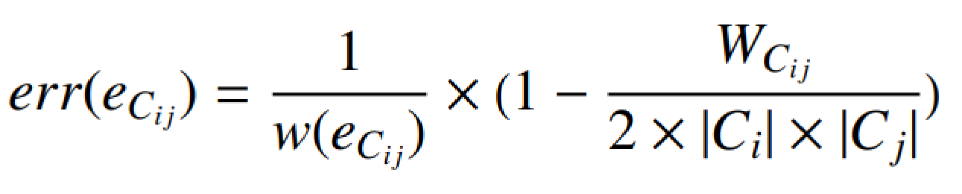
4. ASSIGN ERROR DEGREES

Intra Community Link



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

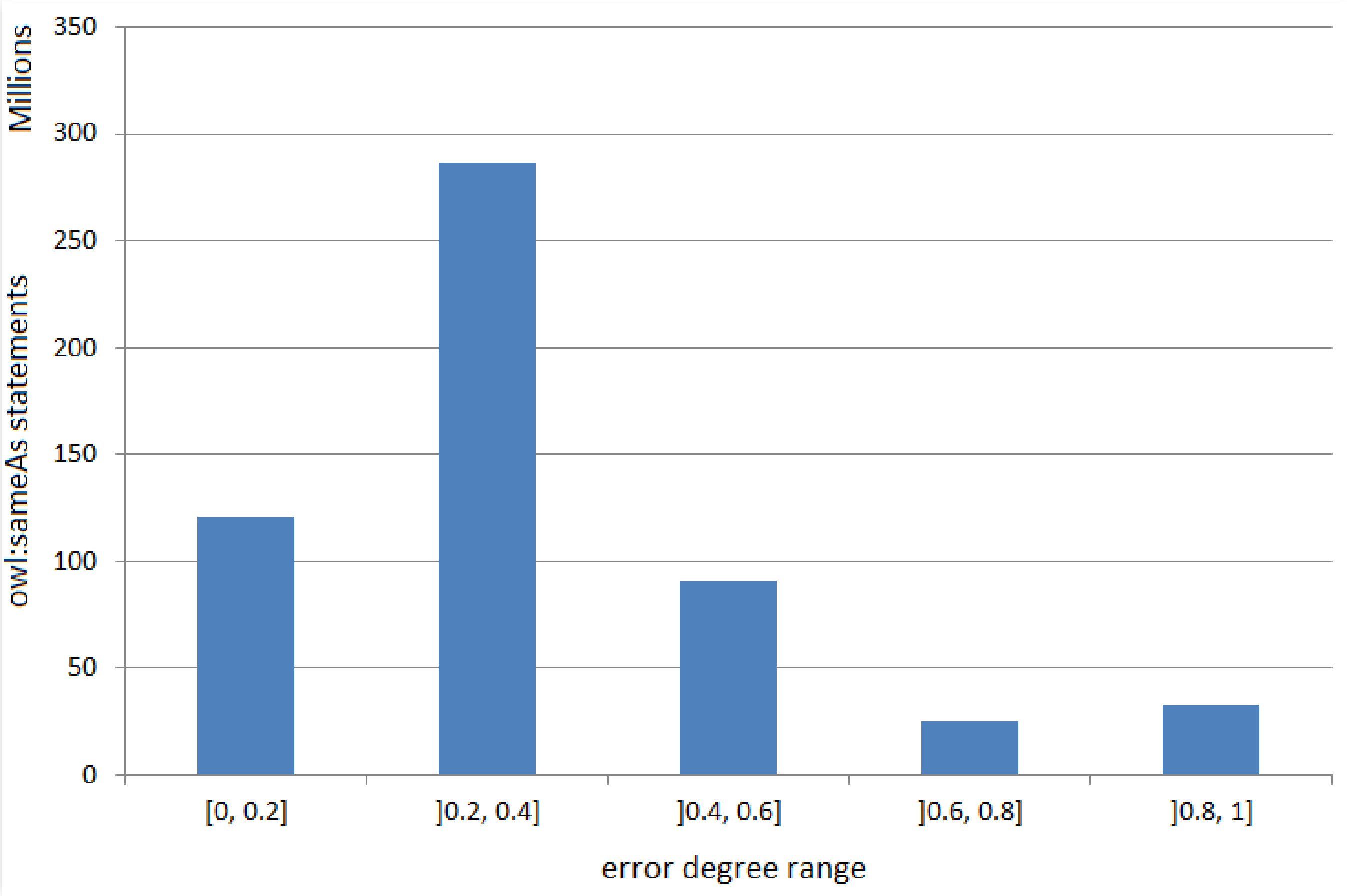
Inter Community Link



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

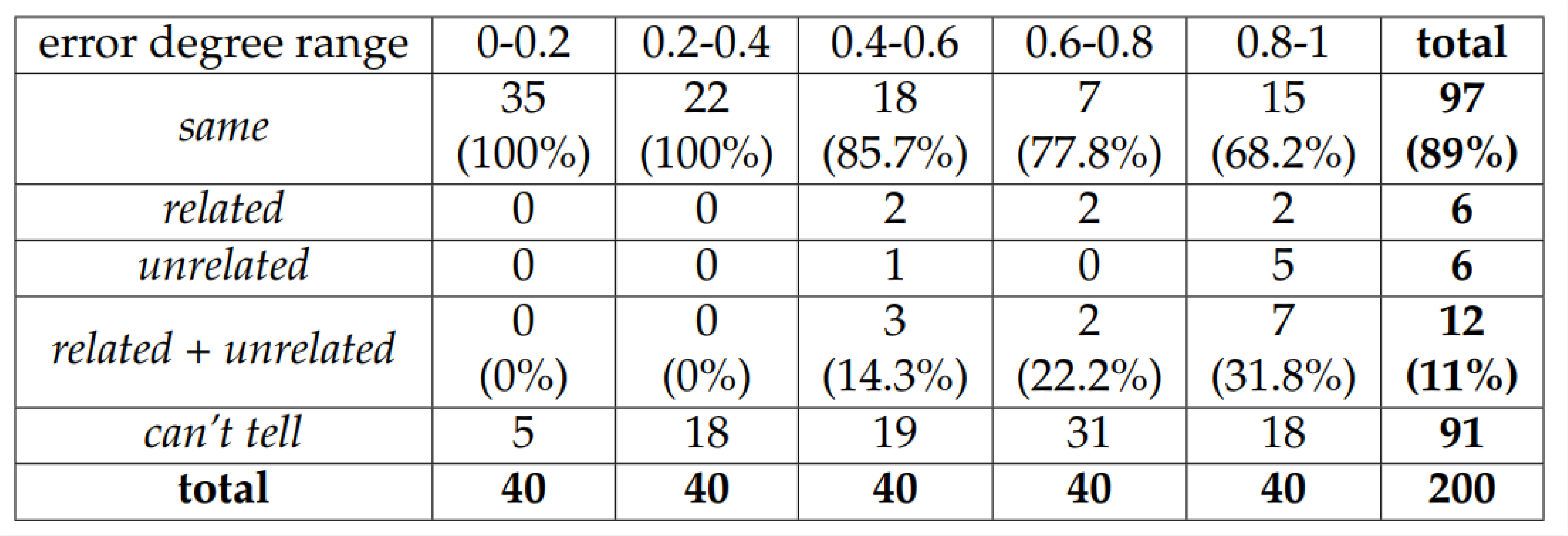
Between 0 and 1 based on the weight of the link and the density of the community(ies)

ERROR DEGREE DISTRIBUTION OF 556M OWL:SAMEAS



EVALUATION

MANUAL EVALUATION OF 200 SAMEAS LINKS

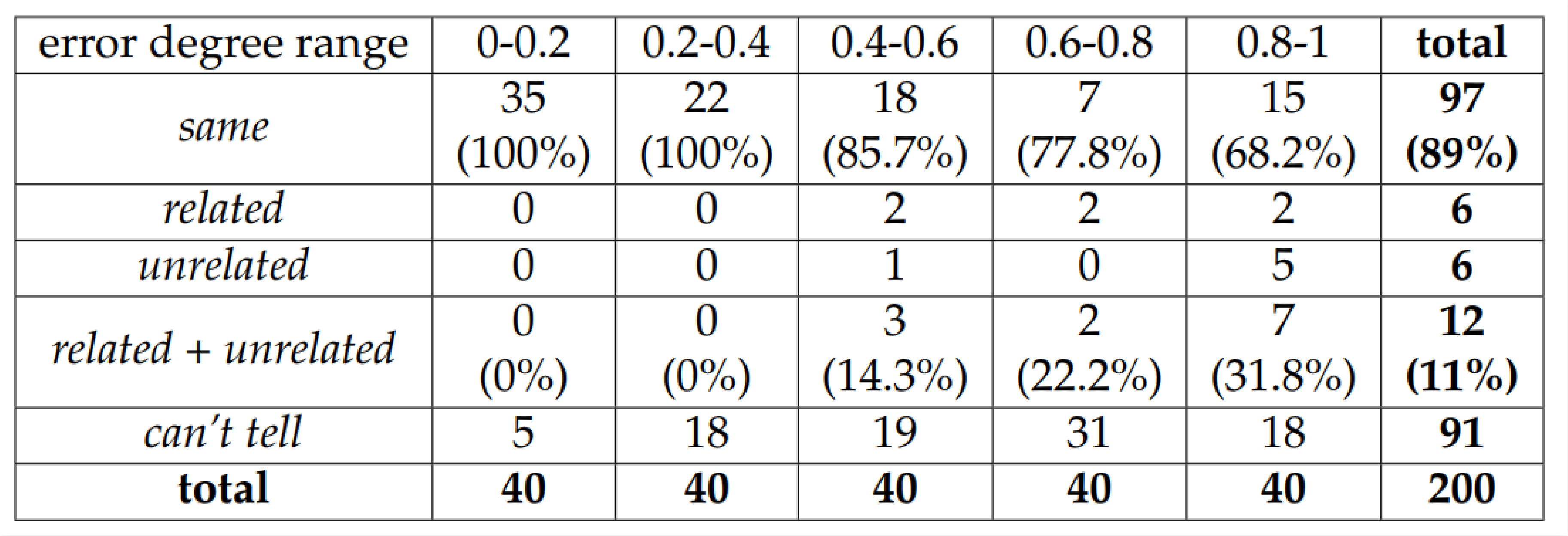


error degree range	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	**total**
same	35 (100%)	22 (100%)	18 (85.7%)	7 (77.8%)	15 (68.2%)	**97** **(89%)**
related	0	0	2	2	2	**6**
unrelated	0	0	1	0	5	**6**
related + unrelated	0 (0%)	0 (0%)	3 (14.3%)	2 (22.2%)	7 (31.8%)	**12** **(11%)**
can't tell	5	18	19	31	18	**91**
total	**40**	**40**	**40**	**40**	**40**	**200**

Result 1. The higher an error degree is, the more likely an owl:sameAs link is erroneous

EVALUATION

MANUAL EVALUATION OF 200 SAMEAS LINKS

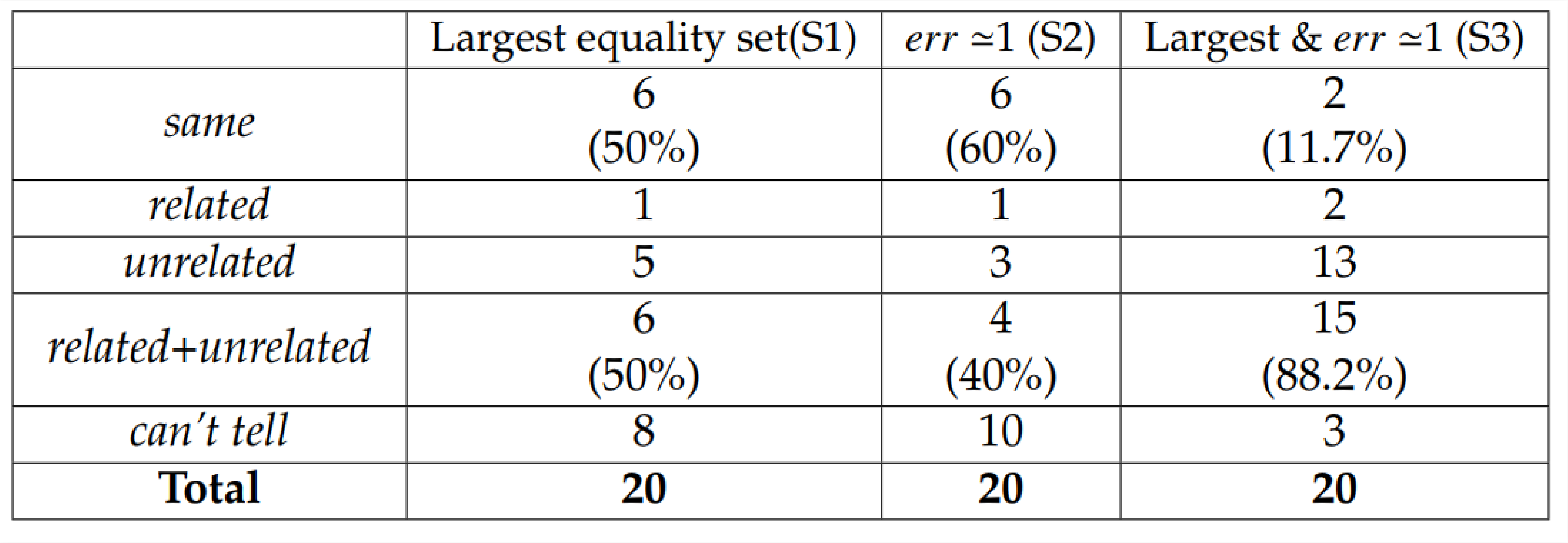


error degree range	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	**total**
same	35 (100%)	22 (100%)	18 (85.7%)	7 (77.8%)	15 (68.2%)	**97** **(89%)**
related	0	0	2	2	2	**6**
unrelated	0	0	1	0	5	**6**
related + unrelated	0 (0%)	0 (0%)	3 (14.3%)	2 (22.2%)	7 (31.8%)	**12** **(11%)**
can't tell	5	18	19	31	18	**91**
total	**40**	**40**	**40**	**40**	**40**	**200**

Result 2. All the evaluated links with an error degree <0.4 are correct

EVALUATION

MANUAL EVALUATION OF 60 SAMEAS WITH ERR >0.9



	Largest equality set(S1)	*err* ≃1 (S2)	Largest & *err* ≃1 (S3)
same	6 (50%)	6 (60%)	2 (11.7%)
related	1	1	2
unrelated	5	3	13
related+unrelated	6 (50%)	4 (40%)	15 (88.2%)
can't tell	8	10	3
Total	**20**	**20**	**20**

Result 3. Links with an err >0.99 and belonging to large equality sets are more likely to be incorrect

EVALUATION - RECALL

We have manually chosen 40 random different terms

(dbr:Facebook, dbr:Strawberry, dbr:Chair)

We made sure there are not explicitly sameAs

(some are in the same equality set)

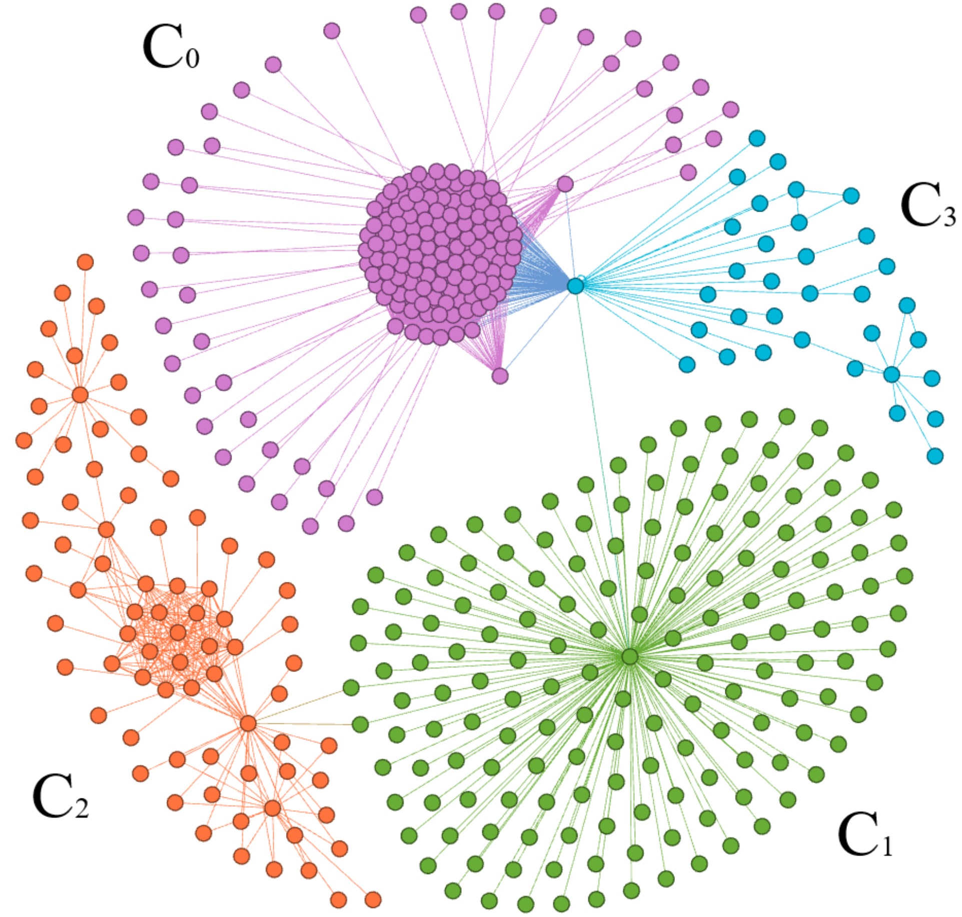
We added all the possible 780 links between them

Result 4. Error degree range from 0.87 to 0.9999.

When the threshold is fixed at 0.99, the recall is 93%

WHO MESSED UP THE LOD?

C0: person; C1: president; C2: government; C3: senator



WHO MESSED UP THE LOD?

```
freebase:m.05b6w1g owl:sameAs dbr:President_Barack_Obama
freebase:m.05b6w1g owl:sameAs dbr:President_Obama
```

```
freebase:m.05b6w1g freebase:type.object.name  "Presidency of B
```

Both owl:sameAs links have are error degree = 0.99999

the only two links in the 'Obama' equality set with err >0.99

CONCLUSION

OUR SOLUTION FOR THE "SAMEAS PROBLEM"

1. Identity Service for the LOD to access:
   - the existing owl:sameAs statements
   - the list of identical terms

2. Detect the incorrect owl:sameAs links in the LOD

IS IT ENOUGH?

Identity is contextual: things can be identical in some contexts and different in other contexts

We need a contextual identity link with formal semantics

J.Raad, N.Pernelle, and F.Saïs
Detection of contextual identity links in a knowledge base, KCap 2017

THANK YOU!

Joe Raad

joe.raad@agroparistech.fr

- J.Raad, W.Beek, F.van Harmelen, N.Pernelle, and F.Saïs
  Detecting Erroneous Identity Links on the Web using Network Metrics, ISWC 2018

- W.Beek, J.Raad, J.Wielemaker, and F.van Harmelen
  sameAs.cc: The Closure of 500M owl:sameAs Statements, ESWC 2018