DETECTING ERRONEOUS IDENTITY LINKS
IN THE WEB OF DATA

Joe Raad\textsuperscript{[1,3]},
Wouter Beek\textsuperscript{[2]}, Frank van Harmelen\textsuperscript{[2]},
Nathalie Pernelle\textsuperscript{[3]}, Fatiha Saïs\textsuperscript{[3]}

joe.raad@agroparistech.fr

\textsuperscript{[1]} INRA, Paris France
\textsuperscript{[2]} VU, Amsterdam The Netherlands
\textsuperscript{[3]} LRI, Orsay France
MOTIVATION
THE WEB NEEDS OWL:sameAs

"Include links to other IRIs, to discover more things."

(4th Linked Data principle)

\[ \langle x, \text{owl:sameAs}, y \rangle \]

means that:

\[ (\forall P)(Px \leftrightarrow Py) \]
"IDENTITY" LINKS IN THE LOD

LOD-a-lot dataset [Fernandez et al. 2017]

<table>
<thead>
<tr>
<th>Link</th>
<th>Number of Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>owl:sameAs</td>
<td>558 M</td>
</tr>
<tr>
<td>skos:exactMatch</td>
<td>566 K</td>
</tr>
<tr>
<td>umbel:isLike</td>
<td>461 K</td>
</tr>
<tr>
<td>skos:closeMatch</td>
<td>371 K</td>
</tr>
</tbody>
</table>
"THE SAMEAS PROBLEM" (1/2)

LOD contains a number of erroneous owl:sameAs

[Ding et al. 2010; Halpin el al. 2010]

- Data Linkage Approaches are not 100% precise
- Lack of formal standardized alternatives
- Concept drift of identity links targets
- Notion of identity is problematic
"THE SAMEAS PROBLEM" (2/2)

The largest identity set contains 177,794 terms that 'should' refer to the same real world entity

However:

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
http://dbpedia.org/resource/Lists_of_christian_religions
...

full list at: https://sameas.cc/term?id=4073
HOW CAN WE DETECT ERRONEOUS SAMEAS LINKS?

Inconsistency-based

Content-based
[Acosta et al. 2013; Paulheim 2014; Cuzzola et al. 2015]

Network Metrics
[Guéret et al. 2012; Sarasua et al. 2017]
REQUIREMENTS

High accuracy, precision and recall
Tested on real world data
Scalable to the Web
No (or minimal) assumptions on the data
  (e.g. UNA, textual description, schema mappings)

No existing approach combines all these criteria
APPROACH & EXPERIMENTS
OVERALL IDEA

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

4 MAIN STEPS
DATASET

LOD-a-lot dataset: 28.3B unique triples
[Fernandez et al. 2017]
1. EXTRACT THE EXPLICIT IDENTITY STATEMENTS

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

\[
\begin{align*}
:a & \text{ owl:sameAs } :b \\
:a & \text{ owl:sameAs } :c \\
:c & \text{ owl:sameAs } :a \\
:d & \text{ owl:sameAs } :e \\
:f & \text{ owl:sameAs } :f
\end{align*}
\]

Eq Set 1

Eq Set 2

OUTPUT: 48.9M Non-Singleton Equality Sets

[Beek et al. 2018]
These 440 identifiers denote the 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
- Low computational complexity
- Outperforms other algorithms

[Blondel et al. 2008; Lancichinetti and Fortunato. 2009; Yang et al. 2016]
4. ASSIGN ERROR DEGREES

Intra-Community Link

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

Inter-Community Link

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

Between 0 and 1 based on the weight of the link and the density of the community(ies)
COMMUNITIES - 'BARACK OBAMA'

C0: person;
C1: president;
C2: government;
C3: senator
EVALUATION
ERROR DEGREE DISTRIBUTION OF 556M OWL:SAMEAS
OBJECTIVES

• Check if links with high error degree have higher probability of incorrectness

• Check what criteria impact mostly results (other than density of communities and symmetry of links)
1. The higher an error degree is, the more likely an owl:sameAs link is erroneous
2. All evaluated links with an error degree ≤ 0.4 are correct

<table>
<thead>
<tr>
<th>error degree range</th>
<th>0-0.2</th>
<th>0.2-0.4</th>
<th>0.4-0.6</th>
<th>0.6-0.8</th>
<th>0.8-1</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>same</td>
<td>35 (100%)</td>
<td>22 (100%)</td>
<td>18 (85.7%)</td>
<td>7 (77.8%)</td>
<td>15 (68.2%)</td>
<td>97 (89%)</td>
</tr>
<tr>
<td>related</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>unrelated</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>related + unrelated</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>3 (14.3%)</td>
<td>2 (22.2%)</td>
<td>7 (31.8%)</td>
<td>12 (11%)</td>
</tr>
<tr>
<td>can't tell</td>
<td>5</td>
<td>18</td>
<td>19</td>
<td>31</td>
<td>18</td>
<td>91</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>200</td>
</tr>
</tbody>
</table>
### Manual Evaluation of 60 SameAs with Err > 0.9

<table>
<thead>
<tr>
<th></th>
<th>Largest equality set(S1)</th>
<th>$\text{err} \approx 1$ (S2)</th>
<th>Largest &amp; $\text{err} \approx 1$ (S3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>same</strong></td>
<td>6 (50%)</td>
<td>6 (60%)</td>
<td>2 (11.7%)</td>
</tr>
<tr>
<td><strong>related</strong></td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>unrelated</strong></td>
<td>5</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td><strong>related+unrelated</strong></td>
<td>6 (50%)</td>
<td>4 (40%)</td>
<td>15 (88.2%)</td>
</tr>
<tr>
<td><strong>can't tell</strong></td>
<td>8</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

3. Links with an err > 0.99 and belonging to large equality sets are more likely to be incorrect.
Erroneous sameAs: $\text{err} > 0.99$ (threshold)

<table>
<thead>
<tr>
<th></th>
<th>TN</th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>97</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>109</td>
</tr>
<tr>
<td>Table 2</td>
<td>6</td>
<td>20</td>
<td>5</td>
<td>8</td>
<td>39</td>
</tr>
<tr>
<td>'Obama' EqSet</td>
<td>30</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>Total</td>
<td>133</td>
<td>22</td>
<td>17</td>
<td>8</td>
<td>180</td>
</tr>
</tbody>
</table>

86% Accuracy  
$$\frac{TN + TP}{Total}$$  

73% Precision  
$$\frac{TP}{TP + FP}$$
Erroneous sameAs: err > 0.99 (threshold)
95 correct DBpedia-Freebase owl:sameAs

78 / 95 links exist in our dataset

77 / 78 links have an err ≤ 0.99
(error degree ranging from 0.52 to 0.94)

98.7% Accuracy

[Acosta et al. 2013]
Manually chose 40 random different terms (e.g. dbr:Facebook, dbr:Strawberry, dbr:Chair)

Made sure they are not explicitly sameAs (but some which are in the same equality set)

Added all the possible 780 links between them

93% Recall
725/780 have an error degree > 0.99 (error degree ranging from 0.87 to 0.9999)
WHO IS MESSING UP THE LOD?

C0: person;
C1: president;
C2: government;
C3: senator
WHO IS MESSING UP THE LOD?

Both owl:sameAs links have an error degree of 0.9999, the only two links in the 'Obama' equality set with err > 0.99.
HOW MESSED UP IS THE LOD?

Our approach can give an approximation based on:

1. nbr of symmetrical sameAs
   (450M with 98% chance of correctness)
2. nbr of non-symmetrical sameAs with err degree ≤0.99
   (105M with a 88.6% chance of correctness)
3. nbr of non-symmetrical sameAs with err degree >0.99
   (1.2M with an erroneous probability between 40 and 88%)

We estimate that 4% of owl:sameAs links in the LOD are erroneous (around 22.5M links)

previous estimations:
2.8% by [Hogan et al. 2012] and 21% by [Halpin et al. 2010]
CONCLUSION & PERSPECTIVES
SUMMARY

Approach for detecting erroneous identity links using the sameAs network's community structure

+ No assumptions on the data (UNA, mapping, desc)
+ Scale to the Web: runtime of 11 hours
+ Accuracy: 86% ; Recall: 93%
- Precision: 40-73% (depending on the Eq Set size)
- Manual Evaluation of only 300 sameAs links
PERSPECTIVES

- Include duplicate owl:sameAs triples
- Consider the Equality Set size in the error degree
- Combine multiple community detection algorithms
- Combine with state-of-the-art approaches
- Replace erroneous owl:sameAs links with contextual identity links [Beek et al. 2016; Raad et al. 2017]
J.Raad, W.Beek, F.van Harmelen, N.Pernelle, and F.Saïs
Detecting Erroneous Identity Links on the Web using Network Metrics.

ISWC 2018, October 8-12, pages 391-407

- Links, Error Degrees, and Equality Sets: https://sameas.cc
- Code: https://github.com/raadjoe/LOD-Community-Detection
- Presentation: https://github.com/raadjoe/LOD-Community-Detection