Detecting Erroneous Identity Links on the Web using Network Metrics

Joe Raad, Wouter Beek, Frank van Harmelen, Nathalie Pernelle, Fatiha Saïs

Detecting Erroneous Identity Links on the Web using Network Metrics

Joe Raad[1,3], Wouter Beek[2], Frank van Harmelen[2], Nathalie Pernelle[3], and Fatiha Saïs[3]

[1] UMR MIA-Paris, INRA, Paris-Saclay University, Paris, France
joe.raad@agroparistech.fr
[2] Dept. of Computer Science, VU University Amsterdam, NL
{w.g.j.beek, frank.van.harmelen}@vu.nl
[3] LRI, Paris Sud University, CNRS 8623, Paris Saclay University, Orsay, France
{nathalie.pernelle, fatiha.sais}@lri.fr

Abstract. In the absence of a central naming authority on the Semantic Web, it is common for different datasets to refer to the same thing by different IRIs. Whenever multiple names are used to denote the same thing, `owl:sameAs` statements are needed in order to link the data and foster reuse. Studies that date back as far as 2009, have observed that the `owl:sameAs` property is sometimes used incorrectly. In this paper, we show how network metrics such as the community structure of the `owl:sameAs` graph can be used in order to detect such possibly erroneous statements. One benefit of the here presented approach is that it can be applied to the network of `owl:sameAs` links itself, and does not rely on any additional knowledge. In order to illustrate its ability to scale, the approach is evaluated on the largest collection of identity links to date, containing over 558M `owl:sameAs` links scraped from the LOD Cloud.

Keywords: Linked Open Data, Identity, owl:sameAs, Communities

1 Introduction

As the Web of Data continues to grow, more and more large datasets – covering a wide range of topics – are being added to the Linked Open Data (LOD) Cloud. It is inevitable that different datasets, most of which are developed independently of one another, will come to describe (aspects of) the same thing, but will do so by referring to that thing with different names. This situation is not accidental: it is a defining characteristic of the (Semantic) Web that there is no central naming authority that is able to enforce a Unique Name Assumption (UNA). As a consequence, identity link detection, i.e., the ability to determine – with a certain degree of confidence – that two different names in fact denote the same thing, is not a mere luxury but is essential for Linked Data to work. Thanks to identity links, datasets that have been constructed independently of one another are still able to make use of each other's information. The most common predicate that is used for interlinking data on the web is the `owl:sameAs`

property. This property denotes a very strict notion of identity that is formalized in model theory. It is defined by Dean et al. [9] as: "an `owl:sameAs` statement indicates that two references actually refer to the same thing". As a result, a statement of the form "x `owl:sameAs` y" indicates that *every* property attributed to x must also be attributed to y, and vice versa.

Over time, an increasing number of studies have shown that `owl:sameAs` is sometimes used incorrectly in practice. For example, Jaffri et al. [15] discuss how erroneous uses of `owl:sameAs` in the linking of DBpedia and DBLP has resulted in several publications being affiliated to incorrect authors. In addition, Ding et al. [10] discuss a number of issues that arise when linking New York Times data to DBpedia. Specifically, they discuss issues that arise when two things are considered the same in some, but not all contexts.

Halpin et al. [13] discuss how the 'sameAs problem', originates from the identity and reference problems in philosophy. In the Semantic Web literature, several approaches have been proposed that focus on limiting this problem. While some approaches consider the introduction of alternative properties that can replace `owl:sameAs` [13], or alternative semantics of the `owl:sameAs` property [1, 22], other approaches focus on the (semi-)automatic detection of potentially incorrect `owl:sameAs` statements [5, 7, 20].

This paper presents a novel approach for the automatic detection of potentially erroneous `owl:sameAs` statements. The approach consists of applying an existing community detection algorithm to an RDF graph that contains solely `owl:sameAs` statements. Based on the communities that are detected, an error degree is calculated for each identity link in the graph. The error degree of an `owl:sameAs` link depends on the density of the community(ies) in which the two terms exist, and whether the identity link is symmetrical or not. It is subsequently used to rank identity links, allowing potentially erroneous links to be identified, and potentially true `owl:sameAs` to be validated.

Since the here presented approach is specifically developed in order to be applied to real-world data, the experiment is run on the largest collection of identity links to date, containing over 558 million `owl:sameAs` links scraped from the LOD Cloud. The evaluation indicates that the calculated error degrees are useful for identifying a large number of correct and erroneous identity links, when applied to this real-world data collection.

The rest of this paper is structured as follows. The next section discusses related work. The approach for detecting potentially erroneous identity links is presented in Section 3. The experiments and the evaluation are described in Section 4, and Section 5 concludes.

2 Related work

This section will give an overview of the related work on detecting erroneous identity links (section 2.1) and existing approaches for community detection

(section 2.2). We will briefly reflect on why we believe community detection to be a particularly good fit for identity error detection in Section 2.3.

2.1 Identity error detection

Source Trustworthiness. An early approach for detecting erroneous identity statements in the Web of Data is idMesh [5], a probabilistic and decentralized framework for entity disambiguation. idMesh hypothesizes that links published by trusted sources (e.g., OpenID-based) are more likely to be correct. The approach detects conflicts between `owl:sameAs` and `owl:differentFrom` assertions by using a graph-based constraint satisfaction solver that exploits the symmetric and transitive nature of the `owl:sameAs` relation. The detected conflicts are resolved based on the iteratively refined trustworthiness of the sources from which the assertions originate.

UNA Violations. Several approaches have made use of the hypothesis that individual datasets apply the Unique Name Assumption (UNA) [7, 23], and that violations of the UNA that are caused by cross-dataset linking are indicative of erroneous identity links. De Melo [7] applies a linear programming relaxation algorithm that seeks to delete the minimal number of `owl:sameAs` statements such that the UNA is no longer violated. Valdestilhas et al. [23] efficiently detect the resources that share the same equivalence class and that belong to the same dataset, and ranks erroneous candidates based on the number of UNA violations.

Content-based. Paulheim [21] represents each identity link as a feature vector in a high dimensional vector space, using direct types and in- and/or outgoing properties. They have tested different outlier detection methods in order to assign a score to each link, indicating the likeliness of being an outlier. Cuzzola et al. [6] propose to calculate a similarity score between the names that are involved in a given `owl:sameAs` link, by using the textual descriptions that are associated to these names (e.g., through the `rdfs:comment` property).

Ontology Axiom Violations. Hogan et al. [14] exploit ten OWL 2 RL rules in order to express the semantics of axioms such as *differentFrom* and *complementOf* in order to detect inconsistencies. Whenever an inconsistent equality set is detected, the erroneous links are identified by incrementally rebuilding the equality set in a manner that preserves consistency. Papaleo et al. [20] exploit class disjointness, (inverse) functional properties, locally complete properties, and property mappings in order to detect inconsistencies in an RDF graph made of the subparts of the two RDF descriptions involved in conflicting statements.

Network Metrics. Finally Gueret et al. [12] hypothesizes that the quality of a link can be determined based on how connected a node is within the network in which it appears. The approach is based on the use of three classic network metrics (clustering coefficient, centrality, and degree), and two Linked Data-specific metrics (`owl:sameAs` chains, and description richness). The approach constructs a local network for a set of selected resources by querying the Web of Data. After measuring the different metrics, each local network is first extended by adding

new edges and then analyzed again. The result of both analyses is compared to the ideal distribution for the different metrics.

2.2 Community detection

Despite the absence of a universally agreed upon definition, communities are typically thought of as groups that have dense connections among their members, but sparse connections with the rest of the network. Community detection is a form of data analysis that seeks to automatically determine the community structure of a complex network [18, 19]. It has applications in statistical physics, mathematics, computer science, biology, and sociology [24]. Importantly, community detection only requires information that is already encoded in the network topology.

Several community detection algorithms exist, as well as several comparative studies. Lancichinetti et al. [16] analyse 12 community detection algorithms by applying them to the LFR graph benchmark [17]. In a more recent study, Yang et al. [24] have evaluated 8 of the most widely used community detection algorithms, again on the LFR benchmark graphs. From both meta studies, the Louvain algorithm emerges as combining a high accuracy with good computational performance.

The Louvain algorithm [4] is a greedy heuristic method, that starts out by assigning a different community to each node of a given network. It then moves each node over to one of its neighbor communities, specifically, neighbors, the one which results in the highest contribution to a modularity score. In the next step, each community from the previous step is regarded as a single node, and the same procedure is repeated until the modularity (which is always computed with respect to the original graph) no longer increases.

2.3 Discussion

We believe community detection to be a particularly good fit for identity error detection, since it can be applied to the network structure of the `owl:sameAs` graph itself. Specifically, the approach that we suggest does not require access to resource descriptions, property mappings, or vocabulary alignments. Also, it does not rely on additional assumptions like the UNA that could be false for some dataset (e.g., datasets that are constructed over a longer period of time and/or by a large group of contributors). Finally, current approaches for identity error detection have not always been applied to real-world `owl:sameAs` links, and no current approach has been evaluated at web scale, i.e., applied to hundreds of millions of links. Since the Louvain algorithm has already been successfully used in other domains, we believe that it can also perform well on the task of detecting `owl:sameAs`-based communities.

3 Approach

This section presents our approach for detecting erroneous identity links by exploiting the community structure of the identity network itself. This section describes the two main steps that our approach is composed of: firstly, the extraction and compaction of the identity network (section 3.1), and secondly, the ranking of each identity link based on the community structure (section 3.2). Algorithm 1 provides an effective procedure for calculating this ranking.

3.1 Identity Network Construction

The first step of our approach consists of extracting the identity network from a given data graph (definition 1).

Definition 1 (Data Graph). *A data graph is a directed and labeled graph* $G = (V, E, \Sigma_E, l_E)$*. V is the set of nodes*[4]*. E is the set of node pairs or edges. Σ_E is the set of edge labels. $l_E : E \to \Sigma_E$ is the mapping from edges to edge labels. $l_E(e)$ denotes the labels of edge e.*

We use e_{ij} to denote the edge between nodes v_i and v_j. From a given data graph G, we can extract the explicit identity network N_{ex} (definition 2), which is a directed labeled graph that only includes those edges whose labels include `owl:sameAs`.

Definition 2 (Explicit Identity Network). *Given a graph* $G = (V, E, \Sigma_E, l_E)$*, the related explicit identity network is the edge-induced subgraph* $G[\{e \in E \mid \{\texttt{owl:sameAs}\} \subseteq l_E(e)\}]$.

We can reduce the size of the explicit identity network N_{ex} into a more concisely represented undirected and weighted identity network I (definition 3), without losing any significant information. Since reflexive `owl:sameAs` statements are implied by the semantics of identity, there is no need to represent them explicitly. In addition, since the symmetric statements e_{ij} and e_{ji} make the same assertion: that v_i and v_j refer to the same thing, we can represent this more efficiently, by including only one undirected edge with a weight of 2. A weight of 1 is assigned for edges which either e_{ij} or e_{ji}, but not both, are present in N_{ex}.

Definition 3 (Identity Network). *The identity network is an undirected labeled graph* $I = (V_I, E_I, \{1, 2\}, w)$*, where V_I is the set of nodes, E_I is the set of edges, $\{1, 2\}$ are the edges labels, and $w : E_I \to \{1, 2\}$ is the labeling function that assigns a weight w_{ij} to each edge e_{ij}. For each explicit identity network $N_{ex} = (V_{ex}, E_{ex})$, the corresponding identity network I is derived as follows:*

- $E_I := \{e_{ij} \in E_{ex} \mid i \neq j\}$
- $V_I := V_{ex}[E_I]$*, i.e., the vertex-induced subgraph.*
- $w(e_{ij}) := \begin{cases} 1, & \text{if } e_{ij} \in E_{ex} \\ 2, & \text{if } e_{ji} \in E_{ex} \end{cases}$

[4] In RDF, nodes are terms that appear in the subject and/or object position of at least one triple.

3.2 Links Ranking

Given $I = (V_I, E_I, \Sigma_{E_I}, w)$, a partitioning of V_I is a collection of non-empty and mutually disjoint subsets $V_k \subseteq V_I$ that together cover V_I. Since the closure of E_I forms an equivalence set (the semantics of the `owl:sameAs` property states that it is reflexive, symmetric, and transitive), it also induces a unique partitioning. We call members of this partition *identity sets*. These partition members correspond to the connected components of I that we call *equality sets* (definition 4).

Definition 4 (Equality Set). *Given an identity network* $I = (V_I, E_I, \{1, 2\}, w)$*, an equality set* Q_k *is a connected component of* I*.*

We want to detect erroneous identity links based on the community structure of each connected component of the identity network. While the number of potential identity links is quadratic in the size of the domain, the representation of equality sets is only linear in terms of the size of the domain. With equality sets, we can implement the following requirements for our algorithm:

- The calculation of erroneous identity links must not have a large memory footprint, since it must be able to scale to very large identity networks, and preferably to all identity statements that appear in the LOD Cloud.
- It must be possible to perform computation in parallel, to allow errors to be detected relatively quickly, preferably directly after the publication of the potential error into the LOD Cloud.
- Calculation must be resilient against incremental updates. Since triples are added to and removed from the LOD Cloud constantly, adding or removing a `owl:sameAs` link must only require a re-ranking of the links within the equality sets that are directly involved in this link.

In order to compute a ranking for the `owl:sameAs` links, we first partition the identity network into different equality sets (several graph partitioning techniques could be applied, such as [2]). Then we detect a set of non-overlapping communities by applying the *Louvain* algorithm [4] for each equality set.

Given an equality set Q_k, the *Louvain* algorithm returns a set of non overlapping communities $C(Q_k) = \{C_1, C_2, \ldots, C_n\}$ where:

- a community C of size $|C|$ (i.e. the number of nodes) is a subgraph of Q_k such that the nodes of C are densely connected (i.e. the modularity of the Q_k is maximized).
- $\bigcup_{1 \leq i \leq n} C_i = Q_k$ and $\forall C_i, C_j \in C(Q_k)$ *s.t.* $i \neq j, C_i \cap C_j = \emptyset$.

We then evaluate each identity link by relying on its weight and the structure of the communities it occurs in. More precisely, to compute the erroneous degree, we distinguish between two types of links: the *intra-community links* and *inter-community links*.

Definition 5. Intra-Community Link. *Given a community* C*, an intra-community link in* C *noted by* e_C *is a weighted edge* e_{ij} *where* v_i *and* $v_j \in C$*. We denote by* E_C *the set of intra-community links in* C*.*

Definition 6. Inter-Community Link. *Given two non overlapping communities C_i and C_j, an inter-community link between C_i and C_j noted by $e_{C_{ij}}$ is an edge e_{ij} where $v_i \in C_i$ and $v_j \in C_j$. We denote by $E_{C_{ij}}$ the set of inter-community link between C_i and C_j.*

For evaluating an *intra-community link*, we rely both on the density of the community containing the edge, and the weight of this edge. The lower the density of this community and the weight of an edge are, the higher the *error degree* will be.

Definition 7. Intra-Community Link Error Degree. *Let e_C be an intra-community link of the community C, the intra-community error degree of e_c denoted by $err(e_C)$ is defined as follows:*

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

where $W_C = \sum_{e_C \in E_C} w(e)$

For evaluating an *inter-community link*, we rely both on the density of the inter-community connections, and the weight of this edge. The less the two communities are connected to each other and the lower the weight of an edge is, the higher the *error degree* will be.

Definition 8. Inter-Community Link Error Degree. *Let $e_{C_{ij}}$ be an inter-community link of the communities C_i and C_j, the inter-community error degree of $e_{C_{ij}}$ denoted by $err(e_{C_{ij}})$ is defined as follows:*

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

where $W_{C_{ij}} = \sum_{e_{C_{ij}} \in E_{C_{ij}}} w(e)$

4 Experiments

4.1 Dataset

We have tested our approach on the LOD-a-lot dataset [11][5], a compressed data file that contains 28 billion unique triples from the 2015 LOD Laundromat Linked Data crawl [3]. This large subset of the LOD Cloud represents our data graph (definition 1).

[5] `http://lod-a-lot.lod.labs.vu.nl`

Algorithm 1: Identity Links Ranking

Input: G: a Data graph
Output: E^{err}: a set of pairs in the from $\{(e_1, err(e_1)), \ldots, (e_m, err(e_m))\}$ with m is the number of edges in the identity network extracted from G

$I_{ex} \leftarrow ExtractSameAsEdges(G)$; // the explicit identity network
$I \leftarrow empty\_graph$; // the identity network
foreach $(e(v_1, v_2) \in I_{ex}$ *and* $v_1 \neq v_2)$ **do**
    if $(I.containsEdge(e(v_2, v_1, 1)))$ **then**
        $I.updateWeight(e(v_2, v_1, 2)$; // set the weight of this edge to 2
    else
        $I.addEdge(e(v_1, v_2, 1))$; // add this edge to I with a weight = 1
$P \leftarrow I.partition()$; // partitioning the graph into equality sets
foreach $(Q \in P)$ **do**
    $C_{set} \leftarrow LouvainCommunityDetectionAlgorithm(Q)$;
    foreach $(e \in C_{set})$ **do**
        if *(e is intra-community-edge(c_i)* **then**
            $err(e) \leftarrow intraCommunityErroneousness(c_i)$;
        else
            // e is an inter-community edge, c_j is the other community to which e is belonging to;
            $err(e) \leftarrow interCommunityErroneousness(c_i, c_j)$;
        $E^{err}.add(e, err(e))$;
return E^{err};

4.2 Quantitative Results

Explicit Identity Network Extraction. We have extracted the explicit identity network (definition 2) from the data graph described above, by performing a Triple Pattern query of the form $\langle ?, \texttt{owl:sameAs}, ? \rangle$ with the HDT C++ library[6]). This returns a stream of distinct identity pairs, as described in [2]. This extraction process takes around four hours using 1 CPU core, resulting in an explicit identity network of 558.9M edges and 179.73M nodes. The explicit identity network is publicly available at `https://sameas.cc/triple`.

Identity Network Construction. From the explicit identify network described above, we build an identity network (definition 3) containing ∼331M weighted edges and 179.67M terms. We leave out ∼2.8M reflexive edges and ∼225M *duplicate* symmetric edges. As a result, we also leave out 67,261 nodes that only appear in such removed edges. This indicates that 68% of the identity network edges are redundantly asserted, with a weight = 2.

Graph Partitioning. The next step consists of partitioning the identity net-

[6] `https://github.com/rdfhdt/hdt-cpp`

work into several equality sets (definition 4). We have deployed an efficient algorithm described in [2] that partitions the identity network into ∼49M equality sets, in just under 5 hours using 2 CPU cores. The identity sets are publicly available at `http://sameas.cc/id`.

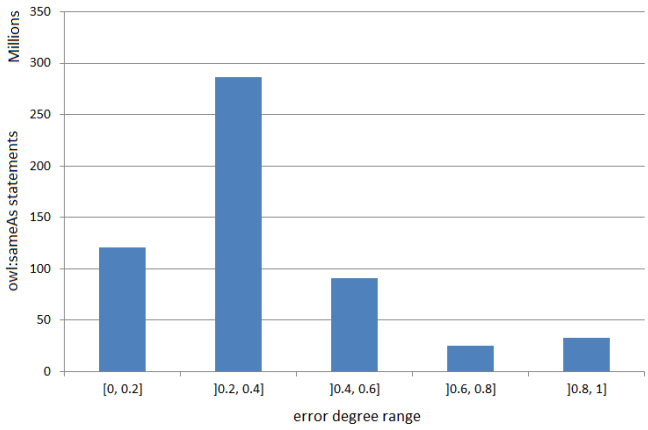

Fig. 1. Error degree distribution of 556M `owl:sameAs` statements

.

Links Ranking. Once the identity network has been partitioned, we apply the *Louvain* algorithm to detect communities in each equality set. We then assign an error degree to all edges of each equality set. This process takes 80 minutes[7], resulting an error degree to each irreflexive[8] `owl:sameAs` statement (∼556M statements) in the explicit identity network. The error degree distribution of these statements is presented in Figure 1, showing that around 73% of the statements have an error degree below 0.4. Whilst this distribution is mainly caused by the high number of symmetrical identity statements in the LOD, it also indicates that most equality sets have a rather dense structure. The 179.67M terms of the identity network were assigned into a total of 24.35M communities, with the communities size varying between 2 and 4,934 terms (averaging ∼7 terms per community). The Java implementation of the link ranking process is available at `http://github.com/raadjoe/LOD-Community-Detection`. The erroneous degree of all the `owl:sameAs` statements are available in our identity web service (`https://sameAs.cc`).

4.3 Community Structure Analysis

In this section we provide a first analysis of the community structure obtained from two equality sets (the largest one and the one about Barack Obama) based

[7] on an 8GB RAM Windows 10 machine, using 2 CPU cores
[8] reflexive statements were discarded in I, and symmetric ones have the same *err*

on the IRIs contained in the communities. In a 2016 study conducted on the same data collection, de Rooij et al. [8] have shown that the social meaning encoded in IRI names significantly coincides with the formal meaning of IRI-denoted resources. Hence, indicating that IRIs can give an idea on the quality of the detected communities.

```
-- Community 258 --  (size = 242)
<http://af.dbpedia.org/resource/Dublin>
<http://am.dbpedia.org/resource/ደብሊን>
<http://an.dbpedia.org/resource/Dublín>
<http://ar.dbpedia.org/resource/دبلن>
<http://ast.dbpedia.org/resource/Ciudá_de_Dublín>
<http://bat-smg.dbpedia.org/resource/Doblėns>
<http://be-x-old.dbpedia.org/resource/Дублін>
<http://br.dbpedia.org/resource/Dulenn>
<http://ca.dbpedia.org/resource/Dublín>
<http://ce.dbpedia.org/resource/Дублин>
<http://commons.dbpedia.org/resource/Dublin_-_Baile_Átha_Cliath>
<http://cs.dbpedia.org/resource/Dublin>
<http://dbpedia.org/resource/Baile_Atha_Cliath>
<http://dbpedia.org/resource/BÁC>
<http://dbpedia.org/resource/Capital_of_Ireland>
<http://dbpedia.org/resource/Capital_of_Republic_of_Ireland>
<http://dbpedia.org/resource/Central_Dublin>
<http://dbpedia.org/resource/City_Center,_Dublin>
<http://dbpedia.org/resource/City_of_Dublin>
<http://dbpedia.org/resource/Dyflin>
<http://dbpedia.org/resource/Europe/Dublin>
<http://dbpedia.org/resource/The_weather_in_Dublin>
<http://dbpedia.org/resource/UN/LOCODE:IEDUB>
<http://dbpedia.org/resource/Visitor_Information_for_Dublin,_Ireland>
<http://dbpedia.org/resource/West_Dublin>
<http://de.dbpedia.org/resource/Dublin>
<http://demo.openlinksw.com/Northwind/Province/ei/Dublin#this>
<http://sws.geonames.org/2964574/>
<http://wordnet.rkbexplorer.com/id/synset-Dublin-noun-1>
<http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Dublin>
```

Fig. 2. Excerpt of the 242 terms included in the community containing the IRI `http://dbpedia.org/resource/dublin`

.

Community Structure in the Largest Equality Set. The largest equality set Q_{max} contains 177,794 terms connected by 2,849,650 undirected and weighted edges. This equality set is the result of the compaction of 5,547,463 distinct `owl:sameAs` statements ($\sim$ 1% of the total number of `owl:sameAs`) and is available at `https://sameas.cc/term?id=4073`. By looking at the IRIs of this equality set, we can observe that it contains a large number of terms denoting different countries, cities, things and persons (e.g. Bolivia, Dublin, Coca-Cola, Albert Einstein, *Literals*, and so on). Clearly showing that this equality set contains many erroneous `owl:sameAs` statements.

Applying the *Louvain* algorithm on Q_{max} resulted in 930 non-overlapping communities, with a size varying from 32 to 2,320 terms per community. As a first interpretation on the community structure, we have solely looked at the IRIs. Despite a few exceptions, we can see that this algorithm is able to group related (and possibly identical) terms in the same community, while keeping

out unrelated terms in other communities. For instance, the community C_{258}, illustrated in Figure 2 contains 242 terms. We can see from this excerpt that most of these terms come from the DBpedia dataset and refer to descriptions of Dublin expressed in different languages: `City of Dublin`, `Capital of Ireland`, `Baile Atha Cliath` (Dublin in Irish), `Dyflin` (the old Norse name for The Kingdom of Dublin), etc. However, we can also see that this community contains terms that do not refer to the city of Dublin, but actually refer to the weather in Dublin or visitor information for Dublin.

With this excerpt of the Dublin community, we can see that an `owl:sameAs` statement between two terms in the same community is not necessarily correct, and requires evaluation as well.

Community Structure in the 'Barack Obama' Equality Set. We present here an analysis of the community structure detected on the equality set Q_{obama} which has a reasonable size and thus easier to analyse. The equality set containing the term `http://dbpedia.org/resource/Barack_Obama` is composed of 440 terms connected by 7,615 undirected and weighted edges. It is built from an explicit identity network of 14,917 `owl:sameAs` statements.

Applying the *Louvain* algorithm on Q_{obama} resulted in 4 non-overlapping communities, with a size varying from 34 to 166 terms per community. This identity set is available at (`https://sameas.cc/term?id=5723`). The resulting community structure of Q_{obama} is presented in Figure 3:

- C_0 **(purple)** includes 166 terms, with 98% of the links of this community representing cross-language symmetrical links between DBpedia IRIs (e.g. `http://fr.dbpedia.org/resource/Barack_Obama`) referring to the person Barack Obama.
- C_1 **(green)** includes 162 terms, mostly DBpedia IRIs of the person Obama in his different roles and political functions (e.g. `http://dbpedia.org/resource/President_barack_obama`, `http://dbpedia.org/resource/senator_obama`).
- C_2 **(orange)** includes 78 terms, mostly referring to the presidency and administration of Barack Obama (e.g. `http://dbpedia.org/resource/Obama_cabinet`, `http://dbpedia.org/resource/Barack_Hussein_Obama_administration`)
- C_3 **(blue)** includes 34 terms from different datasets denoting various entities such as: Barack Obama the person, his senate career, and a misused literal (`"http://dbpedia.org/resource/United_States_Senate_career_of_Barack_Obama`, `"http://dbpedia.org/resource/Barack_Obama"^^xsd:string`).

4.4 Links Ranking Evaluation

In order to evaluate the accuracy of our ranking approach, we have conducted several manual evaluations. The judges relied on the descriptions[9] associated to

[9] the judges were asked to not consider the `owl:sameAs` statements related to the term

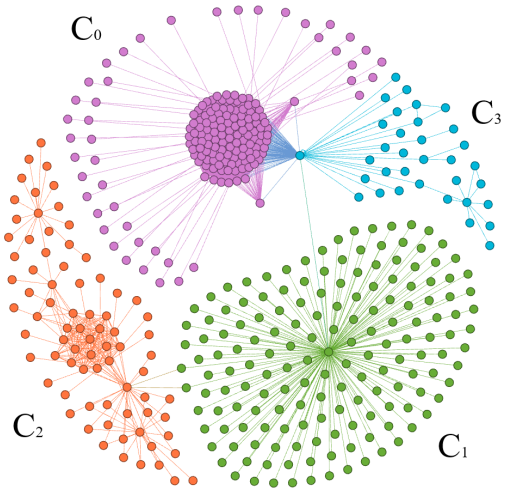

Fig. 3. The communities detected from the equality set containing the term `http://dbpedia.org/resource/Barack_Obama` using the *Louvain* algorithm. The 4 detected communities are distinguished by their nodes' color. The full figure is available at `https://github.com/raadjoe/LOD-Community-Detection/blob/master/Communities-Graph-Obama.svg`.

the terms in the *LOD-a-lot* dataset [11], and did not have any prior knowledge about each link's error degree (i.e. whether they are evaluating a well-ranked link or not). In order to avoid any incoherence between the evaluations, the judges were asked to justify all their evaluations and were given the following instructions: **(a) the same:** if two terms denote the same entity (e.g. Obama and the First Black US President), **(b) related:** not intended to refer to the same entity but closely related (e.g. Obama and the Obama Administration), **(c) unrelated:** not the same nor closely related (e.g. Obama and the Indian Ocean), **(d) can't tell:** in case there are no sufficient descriptions available for determining the meaning of both terms.

A. Accuracy Evaluation in the 'Barack Obama' Equality Set. Firstly, we have relied on the previous observations, made on the community structure presented in Figure 3, to interpret and evaluate the accuracy of our approach:

(i) an `owl:sameAs` statement in C_0 has an average error rate of 0.24. The manual evaluation of 30 random `owl:sameAs` statements shows that they are all true identity links.
(ii) the low density of C_1 has led to several correct `owl:sameAs` statements to have a high error degree (0.9). This is due to the fact that there is only one term linking to all the 161 other terms in this community, with most of these edges being non-symmetrical links.
(iii) the only two `owl:sameAs` statements in this equality set with an error value $\simeq 1$ are the edges in the graph connecting the

IRI `http://rdf.freebase.com/ns/m.05b6w1g` from C_2 to both IRIs `http://dbpedia.org/resource/President_Barack_Obama` and `http://dbpedia.org/resource/President_Obama` from C_1. Relying on their descriptions in the *LOD-a-lot* dataset, we can see that the freebase IRI refers to the presidency of Obama, while the two other IRIs refer to the person Obama, indicating that indeed both statements are incorrect. These two detected incorrect identity statements have led to the false equivalence of the 78 terms of C_2 with the rest of the network's terms.

B. Accuracy Evaluation on a Subset of the Identity Network. In order to evaluate the accuracy over the whole identity network, four of this paper's authors were asked to evaluate a subset of the identity network. The judges were asked to evaluate 200 `owl:sameAs` links (50 links each), representing in an equal manner, each bin of the error degree distribution presented in Figure 1.

Table 1. Evaluation of 200 `owl:sameAs` links, with each 40 links randomly chosen from a certain range of error degree

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**

The results presented in Table 1, shows that the higher an error degree is, the more likely that the link is erroneous. More precisely, we may observe that:

- our error degree is able to identify true `owl:sameAs` links with a high accuracy, since 100% of the evaluated links with an error degree ≤ 0.4. are correct (without considering the "*can't tell*" cases).
- when the error degree is between 0.4 and 0.8, 83.3% of the `owl:sameAs` links are correct. However, in 13.3% of the cases, such links might have been used to refer to two different, but related terms.
- an `owl:sameAs` with an error degree > 0.8 is an unreliable identity statement, referring in 31.8% of the cases to two different, and mostly unrelated terms.

We have further investigated the 22 evaluated identity links with an error degree over 0.8. Two features were observed from the 7 incorrect identity statements: (i) their error degree is most of the times higher than the true `owl:sameAs` links, and (ii) they all belong to equality sets with a higher number of terms than the true ones. To further investigate these observations, we have evaluated 60 additional links with an error degree > 0.9. The first set of links (S1) represents 20 random identity links from the largest equality set. The second set of links (S2) represents 20 random identity links with an error degree $\simeq 1$ (> 0.99). The third

Table 2. Evaluation of 60 `owl:sameAs` links with an error degree > 0.9, with the first set of 20 `owl:sameAs` links (S1) randomly chosen from the largest equality set, (S2) randomly chosen from all links with an error degree $\simeq 1$, (S3) randomly chosen from the largest equality set with an error degree $\simeq 1$

	Largest equality set(S1)	*err* $\simeq 1$ (S2)	Largest & *err* $\simeq 1$ (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**

set of links (S3) represents 20 random links from the largest equality set with an error degree $\simeq 1$. The results presented in Table 2, show that our approach has a high accuracy in detecting erroneous identity links when the threshold is fixed at 0.99 and only equality sets with a high number of terms are considered.

C. Recall Evaluation. In order to calculate the recall of our approach, we have verified how our approach can rank newly introduced erroneous `owl:sameAs` statements. Firstly, we have chosen 40 random terms[10] in the explicit identity network, making sure that all these terms are different, by looking at their descriptions, and that they are not explicitly `owl:sameAs`. From the selected 40 terms, we have generated all the possible 780 undirected edges between them. We added separately, each edge e_{ij} to the identity network with $w(e_{ij})=1$, calculated its error degree, and removed it from the identity network before adding the next one. The error degrees of the newly introduced erroneous identity links range from 0.87 to 0.9999. When the threshold is fixed at 0.99, the recall is 93%.

Results Interpretation. The experiments conducted in this paper, on a subset of 28 billion unique triples of the LOD cloud, shows that there exist many false identity statements on the Web. These erroneous `owl:sameAs` statements have led to the false equivalence of many unrelated terms (e.g. Dublin, Coca-Cola, and Albert Einstein), and many related terms (e.g. Barack Obama the person, and his administration). With a total runtime of 11 hours, these experiments show that an error degree of every available identity link can be computed in practice. Our manual evaluation of these error degrees suggests that:

1. **our approach can validate a large number of identity links in the LOD:** 73% of the identity links have an error degree of [0-0.4]. All the manually evaluated links in this range were judged as true `owl:sameAs` links (100% accuracy, table 1).
2. **our approach can detect numerous erroneous identity links in the LOD:** more than 1.2 million `owl:sameAs` links have an error degree of [0.99-1], with a large number of these links coming from large equality sets (e.g.

[10] we also made sure to include 5 terms that belong to the same equality set

∼13K links in the largest equality set). Up to 88% of the manually evaluated links with these criteria were judged as false identity statements (table 2).
3. **refined content-based approaches are needed** for evaluating the remaining `owl:sameAs` links in the LOD (between 50 and 85% were judged as true identity links).

5 Conclusion

We have presented an approach that aims to detect erroneous `owl:sameAs` statements in RDF data sources. Our approach is uniquely based on the topology of the identity network itself. In order to illustrate its ability to scale, we have evaluated our approach over 558 million `owl:sameAs` statements that are scraped from the LOD Cloud. The evaluation shows that the here introduced calculation of an error degree can indeed be used in order to distinguish between correct and incorrect `owl:sameAs` statements. With a total runtime of 11 hours, these error degrees can be computed in practice. The erroneous degree of all the evaluated `owl:sameAs` statements are available in our identity web service (`https://sameAs.cc`). This will allow others to replicate, check, and hopefully improve upon the here presented results.

The accuracy of the here presented approach could be further improved by combining or comparing results from multiple community detection methods. Since adding a new dataset to the LOD Cloud only requires recalculation of the equivalence sets that are involved in identity assertions within that dataset, it could be useful to test whether the quality of identity links can now be calculated online, e.g., as part of the publication of a dataset into a widely used data catalog.

5.1 Acknowledgment

This work was partially conducted within the MaestroGraph project (612.001.553), funded by the Netherlands Organization for Scientific Research (NWO), and was partially supported by the Center for Data Science, funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

References

1. W. Beek, S. Schlobach, and F. van Harmelen. A contextualised semantics for owl: sameas. In *International Semantic Web Conference*, pages 405–419. Springer, 2016.
2. Wouter Beek, Joe Raad, Jan Wielemaker, and Frank van Harmelen. sameas. cc: The closure of 500m owl: sameas statements. In *European Semantic Web Conference*, pages 65–80. Springer, 2018.
3. Wouter Beek, Laurens Rietveld, and Stefan Schlobach. Lod laundromat (archival package 2016/06). `https://doi.org/10.17026/dans-znh-bcg3`, 2016.
4. V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. of statistical mechanics*, 2008(10):P10008, 2008.
5. Ph. Cudré-Mauroux, P. Haghani, M. Jost, K. Aberer, and H. De Meer. idMesh: graph-based disambiguation of linked data. In *WWW Conf.*, pages 591–600, 2009.

6. John Cuzzola, Ebrahim Bagheri, and Jelena Jovanovic. Filtering inaccurate entity co-references on the linked open data. In *International DEXA Conference*, pages 128–143. Springer, 2015.
7. G. de Melo. Not quite the same: Identity constraints for the web of linked data. In Marie desJardins and Michael L. Littman, editors, *AAAI*. AAAI Press, 2013.
8. S. de Rooij, W. Beek, P. Bloem, F. van Harmelen, and S. Schlobach. Are names meaningful? quantifying social meaning on the semantic web. In *ISWC*, pages 184–199. Springer, 2016.
9. M. Dean, G. Schreiber, S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, and L. Stein. Owl web ontology language reference. *W3C Recommendation February*, 10, 2004.
10. Li Ding, Joshua Shinavier, Tim Finin, and Deborah L McGuinness. owl:sameas and linked data: An empirical study. In *Proceedings of the Second Web Science Conference*, 2010.
11. Javier D. Fernández, Wouter Beek, Miguel A. Martínez-Prieto, and Mario Arias. LOD-a-lot - A Queryable Dump of the LOD Cloud. In *ISWC*, 2017.
12. Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In *Extended Semantic Web Conference*, pages 87–102. Springer, 2012.
13. Harry Halpin, Patrick J Hayes, James P McCusker, Deborah L McGuinness, and Henry S Thompson. When owl: sameas isn't the same: An analysis of identity in linked data. In *ISWC*, pages 305–320. Springer, 2010.
14. Aidan Hogan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres, and Stefan Decker. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10:76–110, 2012.
15. Afraz Jaffri, Hugh Glaser, and Ian Millard. URI disambiguation in the context of Linked Data. In *Linked Data on the Web Workshop (LDOW)*, 2008.
16. Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.
17. Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.
18. Wei Liu, Matteo Pellegrini, and Xiaofan Wang. Detecting communities based on network topology. *Scientific reports*, 4:5739, 2014.
19. Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.
20. Laura Papaleo, Nathalie Pernelle, Fatiha Saïs, and Cyril Dumont. Logical detection of invalid sameas statements in rdf data. In *International Conference EKAW*, pages 373–384. Springer, 2014.
21. Heiko Paulheim. Identifying wrong links between datasets by multi-dimensional outlier detection. In *WoDOOM*, pages 27–38, 2014.
22. Joe Raad, Nathalie Pernelle, and Fatiha Saïs. Detection of contextual identity links in a knowledge base. In *Proceedings of the Knowledge Capture Conference*, page 8. ACM, 2017.
23. André Valdestilhas, Tommaso Soru, and Axel-Cyrille Ngonga Ngomo. Cedal: time-efficient detection of erroneous links in large-scale link repositories. In *International Conference on Web Intelligence*, pages 106–113. ACM, 2017.
24. Z. Yang, R. Algesheimer, and C. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, 6:30750, 2016.