SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl:sameAs in Linked Data

Li Ding, Joshua Shinavier, Zhenning Shangguan, and Deborah L. McGuinness

Tetherless World Constellation, Rensselaer Polytechnic Institute
{dingl, shinaj, shangz, dlm}@cs.rpi.edu

Abstract. Millions of owl:sameAs statements have been published on the Web of Data. Due to its unique role and heavy usage in Linked Data integration, owl:sameAs has become a topic of increasing interest and debate. This paper provides a quantitative analysis of owl:sameAs deployment status and uses these statistics to focus discussion around its usage in Linked Data.

Keywords: owl:sameAs, Linked Data, Network

1 Introduction

The Web of Data is growing rapidly, with an ever-expanding set of inter-connected datasets depicted in the Linking Open Data (LOD) cloud diagram [1]. In the Web of Data, an increasing number of owl:sameAs statements have been published to support merging distributed descriptions of equivalent RDF resources. Although these statements are just binary relations, when all of these owl:sameAs statements are taken together, they form a very large directed graph connecting RDF resources to each other. We will refer to this large graph of RDF resources connected by sameAs statements as a **SameAs network**. SameAs networks are interesting both for their structural properties, e.g. size, diameter and in/out-degree and their semantic properties, e.g. reflexivity, symmetry and transitivity.

According to OWL semantics [2], all RDF resources in a single sameAs network are indistinguishable, such that they can be merged into one RDF resource and change the structure of RDF graph. However, recent literature [3-7], mainly from the Linked Data community, reports many issues related to owl:sameAs usage in the Web of Data: owl:sameAs is often used in ways that do not strictly agree with the official semantics of owl:sameAs in OWL. Some researchers [4, 6] further called for new ontological semantic relations to complement owl:sameAs in capturing similarity relations between RDF resources. To the best of our knowledge, most reported results on owl:sameAs are derived from very small sample datasets, and no statistically significant analysis has been conducted on the deployment status and implications of owl:sameAs in the Web of Data.

We conducted a large scale analysis on SameAs networks extracted from the Web of Data to answer two types of key questions: (i) *How is owl:sameAs actually deployed? How many SameAs networks have been published? Do these SameAs*

networks have interesting topological properties? (ii) *What are the implications of owl:sameAs inference in Linked Data integration? How can owl:sameAs be used to connect the ontologies of the datasets in the LOD cloud?* In order to reduce the bias caused by a small sample dataset, we use the Billion Triple Challenge (BTC) 2010 dataset which covers a significant portion of the Web of Data.

This work provides contributions related to the definition and analysis of SameAs networks. We highlight the practical value of our work in network settings focusing on (1) how Linked Open Datasets are connected and (2) how sameAs networks may be used for automated ontology mapping and error detection. The rest of this paper is structured as follows. Section 2 defines SameAs networks and identifies research problems. Section 3 describes the sample dataset extracted from the BTC 2010 dataset and experiment settings. Sections 4, 5 and 6 report the analytical discoveries on SameAs networks, along with two special classes of networks (Pay-Level Domains and Class-Level Similarity). Section 7 reviews related work. Section 8 concludes our work with future directions.

2 SameAs Networks

The importance of owl:sameAs in Linked Data integration is widely recognized, however there have not been many studies characterizing its usage in very large datasets. The goal of our work was to review existing usage of owl:SameAs in a dataset that contains a significant number of sameAs statements and also to analyze usage in a practical Linked Data integration setting. We therefore will define the notion of a SameAs network and then show a selection of research problems derived from the motivating questions from Section 1.

2.1 Definitions and Notations

Definition 1. **SameAs statement.** A *SameAs statement* is an RDF triple which connects two RDF resources by means of an owl:sameAs predicate.

Definition 2. **Predicate-based Sub-graph Filter.** A *Predicate-based Sub-graph Filter* is a function $H = psf(G, P)$, where H and G are RDF graphs and P is a set of RDF properties. This function returns H which is a sub-graph of G, and the predicate of any triple in H is a member of P.

Definition 3. **SameAs network.** Given an RDF graph G, a *SameAs network SN* in G is a weakly connected component[1] of $psf(G, \{owl:sameAs\})$.

Figure 1 illustrates an example SameAs Network, where an RDF resource "dbpedia:Paul_Allen"[2] is denoted as a node, and a SameAs statement

[1] A weakly connected component is a maximum sub-graph where all pairs of nodes are by an undirected path. See http://mathworld.wolfram.com/WeaklyConnectedComponent.html

"dbpedia:Paul_Allen owl:SameAs umbel:Paul_Allen" is denoted as a directed arc. This figure also exhibits additional interesting structural patterns: (i) two RDF resources could be linked by one-way or and reciprocal owl:sameAs statements; and (ii) there exist authority nodes (with high in-degree, e.g. dbpedia:Paul_Allen) and hub nodes (with high out-degree, e.g. freebase:guid.9202a8c04000641f800000000002e633).

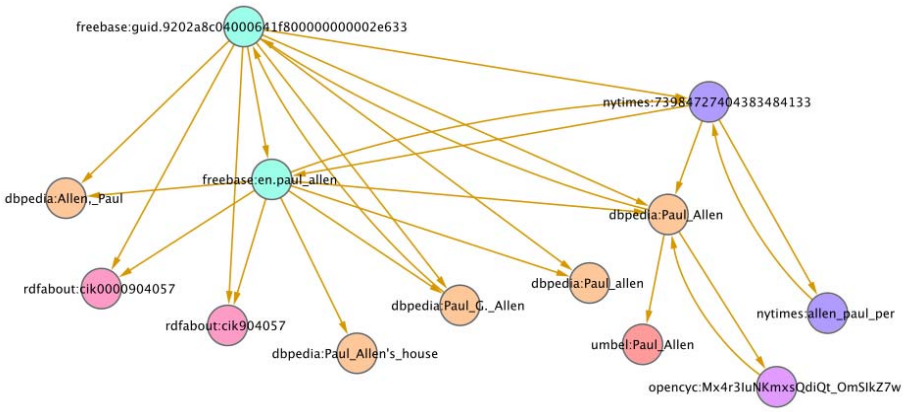

Figure 1. An example SameAs network about "Paul Allen" .

The official semantics of owl:SameAs is specified in OWL [2]: *"an owl:sameAs statement indicates that two URI references actually refer to the same thing."* Recent studies reported diverse usage that is NOT consistent with the official semantics:

- Is owl:sameAs symmetric? Vatant [7] suggested that owl:sameAs, when used in mashup, is not necessarily a symmetric property, i.e., "X owl:sameAs Y" does not imply "Y owl:sameAs X". Therefore, two RDF resources X and Y are considered to be strongly equivalent only when their owners make reciprocal SameAs statements.
- Is owl:sameAs transitive? Jaffri et al [5] reported that the equivalence relationship represented by owl:sameAs is often context-dependent, and is accurate only within the context of particular applications. While transitivity is automatically granted by OWL semantics, SameAs statements asserted in the Web of Data seldom guarantee transitivity.

2.2 SameAs Networks Analysis

In order to analyze the deployment status and implications of SameAs Networks, we identify the following three research problems:

How have SameAs Networks been deployed on the Web of Data? Since we are not the owners of the SameAs statements in the Web of Data, it would be quite subjective

[2] Throughout this paper, we use QName to encode URI reference and Turtle to encode RDF triples and RDF graphs. See http://www.w3.org/TeamSubmission/turtle/.

to speculate the intended semantics of owl:sameAs. In order to produce objective and convincible reports, we focus on the structural properties of SameAs networks. In order to avoid the bias caused by small sample datasets, we collected a large sample dataset from the real world Web of Data. Section 3 provides a quantitative analysis of the dataset.

What are the common interests among Linked Data publishers? Since there are many URIs using "dbpedia" for a namespace in the example SameAs network in Figure 1, it is possible to summarize SameAs statements to higher level connections to provide an overview of SameAs networks. We are particularly interested in "pay-level domain" (PLD)[3] as it is frequently used to identify Linked Data publishers and can often be connected to LOD datasets via one-to-one mappings. Now, with such summarization, users can analyze how and why Linked Data publishers (or LOD datasets) are inter-connected via SameAs statements.

How will Web ontologies be affected by owl:sameAs inference? Mapping Web ontologies is a well-known difficult problem due to the high cost of manually asserting mapping relations among ontological terms. Instance-based approaches have been used in mapping RDFS/OWL classes, i.e. two classes are considered "associated" if they share common instances. Now, with owl:sameAs inference, users may merge different RDF resources and thus find more associated classes.

3 Building ESameNet Dataset and Experiment Settings

In order to study the three problems identified in section 2.2, we will extend SameAs networks with additional information:

- *PLD statements*, each RDF resource can be connected to a literal name identifying a PLD. PLD statements can be pre-computed before the creation of SameAs networks and stored in triples using ex:hasPLD as predicate.
- *Type statements*, each RDF is connected to zero-to-many RDFS/OWL classes via rdf:type. Type statements are already asserted in the RDF graph from which SameAs networks were obtained.

Definition 4: **Extended SameAs network**. Given an RDF graph *G*, an *extended SameAs network ESN* is constructed by extending a SameAs network *SN* of *G* with additional nodes and arcs. Besides the RES world, i.e. the world of all RDF resources in *SN*, two more worlds of nodes will be added: (i) the CLS world, i.e. a world of RDFS/OWL classes; (iii) the PLD world, i.e. a world of PLD names. A new node *n* will be added when there exists a PLD (or Type) statement *s* that links from a node in *SN* to *n*. Meanwhile, the corresponding statement *s* will be added as a new arc.

[3] A PLD is an internet domain that requires payment at a generic top-level domain (gTLD) or country code top-level domain (ccTLD) registrar. PLDs are usually one level below the corresponding gTLD (e.g., dbpedia.org vs. org), with certain exceptions for cc-TLDs (e.g., ebay.co.uk, det.wa.edu.au) [8].

Figure 2 illustrates an example fragment of an extended SameAs network, including: RDF resources, e.g. *dbpedia:Virginia* and *fbase:en.virginia*; PLD statements, e.g. "*dbpedia:Virginia* ex:hasPLD "*dbpedia.org.*" and CLS statements, e.g. "*dbpedia:Virginia* rdf:type yago:StatesOfTheUnitedStates, dbpedia-owl:Place."

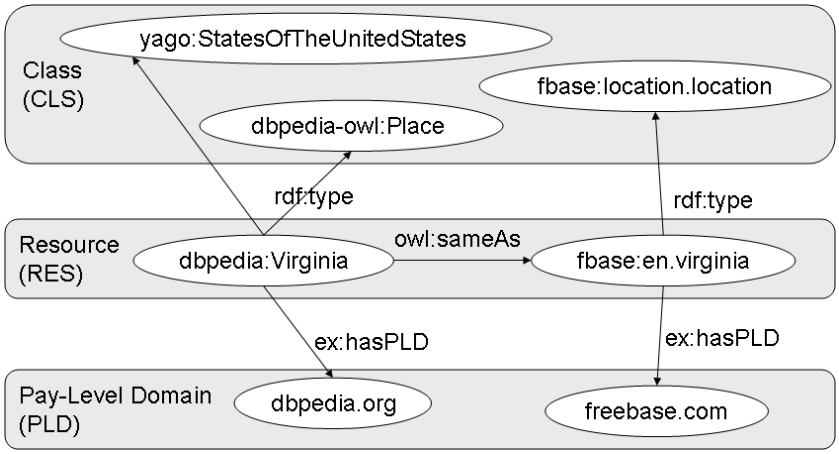

Figure 2. An example fragment of an extended SameAs network .

Our study is based on the **ESameNet** dataset (a collection of extended SameAs networks) extracted from the BTC 2010 dataset. We chose this dataset for two reasons: (i) With approximately 9 million SameAs statements, it constitutes a large-scale sample dataset which is suitable for providing statistical results with high confidence; (ii) Since the BTC 2010 dataset was gathered by crawling the Web based on seeding datasets provided by major Semantic Web search engines, it can be considered as a representative sample of the Web of Data, with relatively low sample distribution bias.

The ESameNet dataset is publicly available[4] in N-Quads[5] format and it consists of the following three subsets:

- **SameAs statements**. We copied all SameAs statements in the BTC 2010 dataset and removed invalid and duplicate statements. A few SameAs statements do not comply with Definition 1 (SameAs statement), e.g. some simply connect an RDF resource to a literal string[6]. From 9,358,227 valid SameAs statements, we obtained 8,711,398 triples after removing duplications. These statements covered 6,932,678 unique RDF resources with URI (aka. URI resource) and 645,400 blank nodes.
- **Type statements**. We copied all rdf:type statements for RDF resources mentioned in BTC 2010 dataset and found 552,622,105 such statements. These statements covered 488,138,983 distinct RDF resources, and 168,503 distinct RDFS/OWL classes.

[4] See http://tw.rpi.edu/2010/ESameNet
[5] http://sw.deri.org/2008/07/n-quads/
[6] E.g. <http://sw.nokia.com/language-1/zh-CH> owl:sameAs "zh_CH"^^xsd:lang.

- **PLD statements**: We extracted PLD (pay-level domain) statements by parsing the URI of RDF resources in SameAs networks using regular expression. For RDF resource with HTTP URI, we can directly extract its PLD and create the PLD statement. For blank nodes (or RDF resources with non-HTTP URI), we assume they share the same PLDs as the named graphs which host the corresponding SameAs statements. These statements covered 967 distinct PLDs.

In our experiments, we used the AllegroGraph triple store (version 4.0)[7] and the Allegro Common Lisp (version 8.2)[8] programming environment to load the entire BTC 2010 dataset and extract the ESameNet dataset. All of the computational tasks described in this paper were executed on a server with 2x Quad-Core Intel Xeon CPU 2.33GHz CPU, 64GB physical memory and 4 TB hard disk space.

4. Basic Properties of SameAs Networks

We first analyze the basic properties of SameAs Networks in the ESameNet dataset. Each SameAs network is essentially a graph of URIs connected by non-redundant owl:sameAs statements. Due to the difficulties and limitations of automatic entity resolution, the creation of owl:sameAs statements is usually costly and requires manual efforts. Therefore, there are fewer owl:sameAs statements in the Web of Data than one might expect.

Weakly connected components. Overall, the ESameNet dataset contains 6,932,678 URI resources connected by 8,711,398 unique owl:sameAs statements. The graph consists of 2,890,027 weakly connected components, each of which covers on average 2.4 URI resources. The average path length of the graph is only 1.07, which is consistent with this very small average component size (see Figure 3); most components are simply pairs of nodes joined by (usually reciprocal) owl:sameAs links. There are a small number of larger components, including 41 components with hundreds of resources, and two components with thousands of resources. This observation implies that the typical size of SameAs networks is either a small constant or growing slowly; therefore, performing transitive inference on individual SameAs networks is not expensive and could be parallelized. A manual inspection of individual components revealed that the vast majority were star-like in structure, consisting of a single central resource connected to a number of peripheral resources. SameAs networks are not large and complex networks like those of foaf:knows, or even shallow tree-like structures like those of rdfs:subClassOf. Furthermore, SameAs networks tends to have small size components: 24,559 persons were found in the largest component of the foaf:knows network in 2005 [9] vs. 5000 resources were found in the largest component component in SameAs networks in 2010. One potential explanation could be that Linked Data principles are in favor of reusing URIs rather than duplicating resource decriptions in many distinct LOD datasets. An

[7] http://www.franz.com/agraph/allegrograph/
[8] http://www.franz.com/products/allegrocl/

alternative explanation is that people simply haven't done enough large-scale linking yet[9] due to technology limitations.

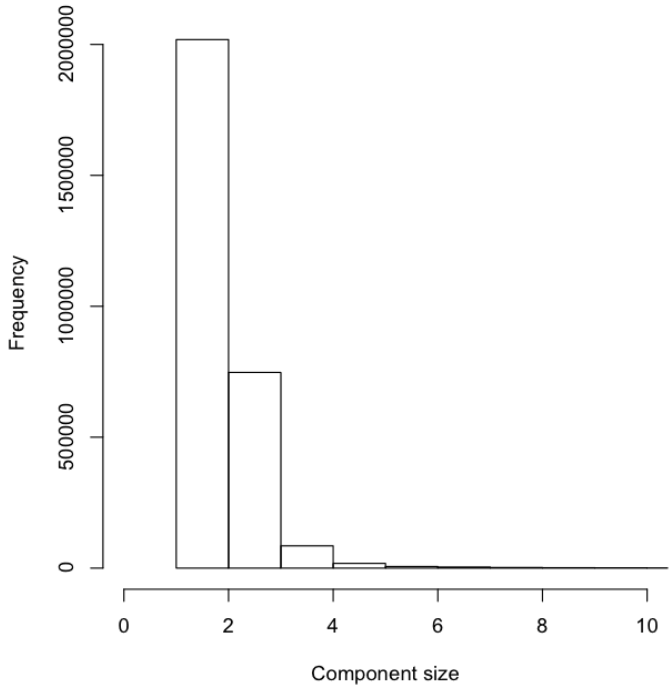

Figure 3. Histogram of the size of SameAs Networks in the ESameNet dataset .

Degree distribution. We investigated the overall in-degree distribution of ESameNet as it measures the popularity (or authority) of resources in sameAs networks[10]**.** Having plotted the in-degree distribution on a log-log scale, we can see that it exhibits the power law pattern characteristic of scale-free networks. We also noticed that there are slightly more resources with an owl:sameAs in-degree of 1 (that is, 2,974,914 resources) than one would expect of a power law distribution (see Figure 4), and there are also slightly more resources in the 10 to 20 range of in-degree than one would expect. The resources at the high end of the distribution contain on the order of 4,000 inbound owl:sameAs links. Note that we omitted resources with no inbound links.

[9] This alternative explanation is kindly suggested by reviewers of this paper.

[10] We skipped out-degree analysis to save space. The out-degree is typically controlled by the publishers for sameAs statements, but the in-degree shows how many publishers are willing to link to a resource using owl:sameAs.

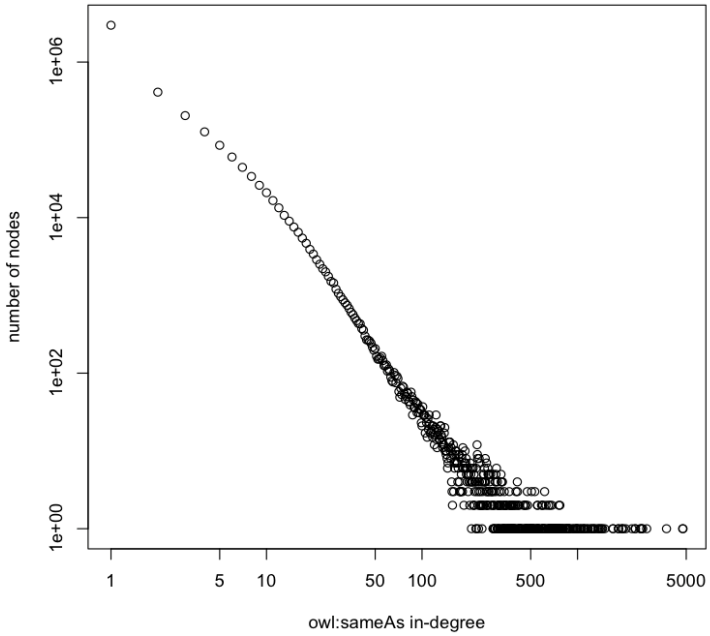

Figure 4. The in-degree distribution of RDF resources in ESameNet .

5. Pay-Level-Domain (PLD) Network Analysis

In order to gain deeper understanding of the common interests between different Linked Data publishers, users may demand a high-level meaningful network to abstract the SameAs networks. The PLD statements provide an ideal opportunity to meet this demand because a PLD can often be used to identify Linked Data publishers and millions of RDF resources in ESameNet can be reduced to hundreds of PLDs.

5.1 Definitions and Notations

Definition 5. **PLD network**. A *PLD network* is a weighted directed graph where (i) each node denotes a unique PLD (labeled by PLD name); (ii) each arc links two PLDs. The weight of an arc <*pld1*, *pld2*> is calculated by counting the unique SameAs statements between any possible pair of *u1* and *u2*, where (*u1* ex:hasPLD *pld1*) and (*u2* ex:hasPLD *pld2*), normalized by the out degree of *pld1*.

Intuitively, the PLD network is an abstraction of SameAs Networks where each PLD groups some RDF resources. Arcs in PLD network are created using the following SPARQL query:

```
SELECT ?pld1  ?pld2
WHERE { ?u1  ex:hasPLD ?pld1 .  ?u2  ex:hasPLD ?pld2 . ?u1 owl:sameAs? u2 . }
```

Figure 5 shows the largest (also the most interesting) cluster in the PLD network[11] generated from the ESameNet dataset, plotted using Cytoscape [10]. In this diagram, the size of a node is determined by the sum of the weights of both its incoming and outgoing arcs. The thickness of an arc is determined only by its weight. For the purpose of visual clarity, we omit arcs whose weight is less than a threshold (0.00001 in this study with 0.069 being the maximum weight), and self-loops (arcs linking from a node to the node itself). The color of a node is randomly assigned, with the guarantee that no two nodes have the same color. We adopt the "Organic" graph layout provided by Cytoscape to render this diagram to visually highlight clusters.

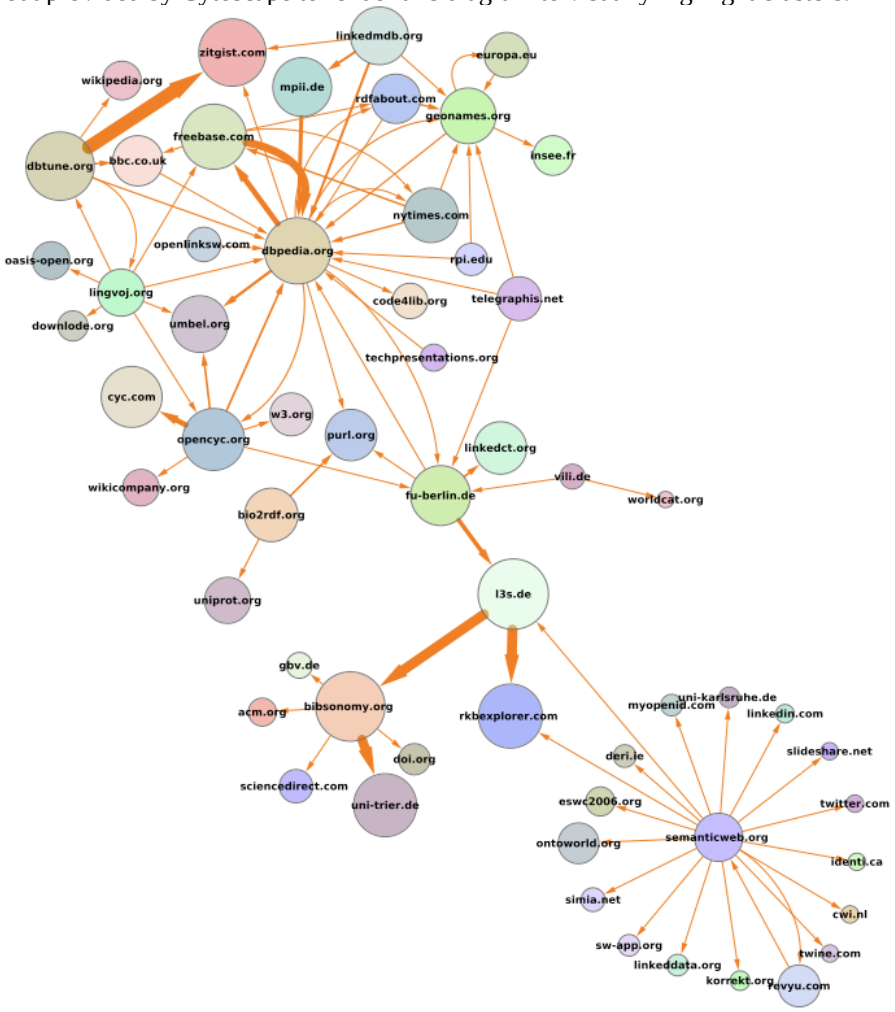

Figure 5. The largest cluster in the PLD Network generated from ESameNet.

[11] Due to space limitations, only the most significant cluster is shown. Other clusters can also be generated using the same method and tools as discussed.

5.2 Implications of the PLD Network

The PLD network is an abstraction of SameAs networks in that it establishes connections at PLD level based on instance-level SameAs statements, while retaining the basic structure of one-per-dataset nodes connected by links in a star-like fashion. It can help us gain better insights to the following research problems:

How are data publishers connected? The PLD network provides intuitive and straightforward insights into how publishers are connected via owl:sameAs assertions and what communities are potentially emerging. Figure 5 shows a clear depiction of the associations between different data publishers, in which thicker arcs reflect intensive occurrences of owl:sameAs assertions between corresponding domains. By using appropriate clustering algorithms which apply to any generic graph (e.g., social network), communities of data publishers can be easily identified by eyes. Nodes inside such a cluster can be considered as covering similar topics from possibly different perspectives. In Figure 5, some clusters can be visually identified. The cluster with the set of PLD nodes {*ls3.de, rkbexplorer.com, uni-tier.de, sciencedirect.com, acm.org, gbv.de, doi.org*} represents a community whose members publish data about scientific publications. Other clusters centering on bioinformatics and Semantic Web communities can also be easily identified. In general, we believe that applying novel clustering algorithms to this large-scale PLD network will facilitate detection of communities that share common knowledge and interests. We perceive this as an interesting future research direction.

Why are data publishers connected? After determining which Linked Data publishers are connected via owl:sameAs assertions, it is natural to further investigate why they are connected. Although it is possible to achieve this goal by manually analyzing Figure 5, it usually requires strong expertise in Linked Data, and thus is labor-intensive and error-prone. With the help of rdf:type information, semi-automatic or even automatic ways of explaining such connectivity is possible.

For all owl:sameAs statements between the source PLD *d1* and target PLD *d2*, we can retrieve the rdf:type information for *u* and *v* using the following SPARQL query:

```
SELECT ?subj_type ?obj_type
WHERE {
        ?s ex:hasPLD “d1”.
        ?o ex:hasPLD “d2”.
        ?s a ?sub_type.
        ?o a ?obj_type.
}
```

Then comparing the *k*-most frequently used types in *d1* with the *k*-most frequently used types in *d2* can help us to understand how the instance resources served by *d1* and *d2* are connected. Table 1 lists the top five (if exists) type labels for the source and target PLD of arcs.

Arc	Top-5 Types in Source PLD	Top-5 Types in Target PLD
<dbtune.org, zitgist.com>	rdfs:Resource: 2864 mo:Track: 2382 mo:Record: 280 mo:MusicArtist: 202	mo:MusicArtist: 99515 mo:MusicGroup: 61368 foaf:Group: 61368 mo:Record: 58245 mo:SoloMusicArtist: 26058
<l3s.de, bibsonomy.org>	foaf:Document: 366416 swrc:InProceedings: 254905 swrc:Article: 104295 swrc:Proceedings: 4164 swrc:Book: 550	N/A
<l3s.de, rkbexplorer.com>	foaf:Document: 366073 swrc:InProceedings: 254567 swrc:Article: 104294 swrc:Proceedings: 4161 swrc:Book: 549	N/A
<bibsonomy.org, uni-trier.de>	swrc:InProceedings: 308486 swrc:Article: 13339 swrc:Proceedings: 3216 swrc:InCollection: 1284 owl:Ontology: 89	N/A
<freebase.com, dbpedia.org>	freebase:base.intellectualproperty.valuable_item: 240685 freebase:medicine.hospital: 51587 freebase:user.morrowjtm.default_domain.sexuality: 46726 freebase:base.onlineadvertising.ad_pricing_model: 24968 freebase:user.ericqianli.default_domain.css: 24123	yago:NeighbourhoodsOfLewisham: 4312 RailwayStationsInLewisham: 638 dbpedia-owl:ProtectedArea: 564 yago:HighSchoolsInCentralPennsylvania: 524 yago:IndigenousPeoplesOfEurope: 519

Table 1. Top five most frequently used types for each arc in Table 1.

The first row in Table 1 indicates that both PLD *d1* = dbtune.org and PLD *d2* = zitgist.com are publishing data about music, because the top five types related to all owl:sameAs links between them are generally well aligned and are using concepts from the Music Ontology[12]. Row 2, 3, and 4 are all missing the type information in the target PLD, which indicates that cross-PLD owl:sameAs links do not have type information for resources in the target PLD. Finally, the top five types in the source and target PLD do not align very well in the last row. This might be due to the vast amount of general human knowledge encoded by dbpedia.org and freebase.com, as well as the unique role of "knowledge hubs" that they have been playing on the Web. Actually, the top-*k* types discussed here can also be used to form a more complete view of either the source or the target PLD, in which case the owl:sameAs statements function as a clue for discovering more information for either side.

[12] Music Ontology: http://musicontology.com/

6. CLS Network Analysis

In order to show how Web ontologies are affected by owl:sameAs inference, we select an ontology mapping use-case: detecting the relations between two RDFS/OWL classes. Two classes are considered overlapping when they share common instances. Classes inter-connected by such "class-overlap" relation form a Class-Level Similarity (CLS) network. With the CLS network, users can automatically detect clusters of classes and ontology mappings using machine learning techniques.

6.1 Definitions and Notation

Definition 6: **CLS network**. A *CLS network* is a weighted directed graph of classes where (i) each node denotes a unique RDFS/OWL class; (ii) each arc links two classes using one of the following relations: equivalence, subclass-of, disjointness and class-overlap. While the first three types of relations can be mapped to OWL properties, the last one cannot. The weight of an arc is calculated based on the number of common instances shared by the two classes linked by the arc.

As shown in Table 2, A CLS network can be constructed using SPARQL queries, namely Query A and Query B. Note that Query B leverages owl:sameAs inference to derive additional class-overlap relations, and it simply assumes that owl:sameAs is neither symmetric nor transitive. Other possible assumptions are left for future study.

Query A	CONSTRUCT ?C1 ex:overlaps ?C2 WHERE { ?s a ?C1, ?C2. filter (?C1!=?C2) }
Query B	CONSTRUCT ?C1 ex:overlaps ?C2 WHERE { ?u1 a ?C1 . ?u2 a ?C2 . ?u1 owl:sameAs ?u2. filter (?C1 != ?C2) }

Table 2. Two SPARQL queries for generating class-overlap relations.

6.2 CLS Network and Enhancement

We executed Query A on all Type statements in ESameNet to build a CLS network CLS-ALL, which contains 168,503 unique nodes (RDFS/OWL classes) and hundreds of millions of arcs. Overall, the in-degree of classes (i.e. how many instances the classes have) follows a power-law distribution: about 45% (77 K) classes only have one instance, while a few have over 100 million instances each.

Focusing on the RDF resources connected by SameAs statements, we created a smaller CLS network CLS-SAME, which contains 6,555 unique nodes (RDFS/OWL classes) and 21,628 arcs (weighted differently) using Query B. Although CLS-SAME is much smaller than CLS-ALL, it helps users to quickly gather additional pairs of

classes for determining class-level relations. Table 3 lists 20 class pairs in the CLS-SAME dataset. We found a couple of interesting observations:

- The rows with type [EQ] show that some class pairs could be mapped via equivalence relation because their URIs have the same local-name. This kind of class pairing can be used to guess equivalence relations.
- The rows with type [ERR] show that some class pairs may also be inappropriate mappings after checking their ontological definitions. Although this kind of class pairing does not help ontology mapping, it does help users to detect potential errors in Linked Data.
- The rows without a type label show that it is hard to determine the mapping relations between the class pairs by checking their URIs or ontological definitions. This kind of case usually involves a general-purpose class, such as <http://semantic-mediawiki.org/swivt/1.0#Subject>. This kind of class paring may be used to guess sub-class relations.

type	FROM	TO
	<http://semantic-mediawiki.org/swivt/1.0#Subject>	<http://xmlns.com/foaf/0.1/Person>
EQ	<http://www.w3.org/2002/07/owl#Class>	<http://www.w3.org/2000/01/rdf-schema#Class>
ERR	<http://www.w3.org/2002/07/owl#Class>	<http://www.w3.org/2002/07/owl#Thing>
	<http://www.geonames.org/ontology#Code>	<http://www.w3.org/2004/02/skos/core#Concept>
	<http://www.w3.org/2004/02/skos/core#Concept>	<http://www.geonames.org/ontology#Code>
EQ	<http://www.daml.org/2001/09/countries/iso-3166-ont#Country>	<http://rdf.geospecies.org/ont/geospecies#Country>
EQ	<http://www.geonames.org/ontology#Country>	<http://rdf.geospecies.org/ont/geospecies#Country>
	<http://semantic-mediawiki.org/swivt/1.0#Subject>	<http://referata.com/wiki/Special:URIResolver/Category-3APeople>
	<http://referata.com/wiki/Special:URIResolver/Category-3APeople>	<http://semantic-mediawiki.org/swivt/1.0#Subject>
	<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property>	<http://www.w3.org/2002/07/owl#ObjectProperty>
	<http://semantic-mediawiki.org/swivt/1.0#Subject>	<http://xmlns.com/foaf/0.1/Agent>
	<http://semantic-mediawiki.org/swivt/1.0#Subject>	<http://discoursedb.org/wiki/Special:URIResolver/Category-3APositions>
	<http://discoursedb.org/wiki/Special:URIResolver/Category-3APositions>	<http://semantic-mediawiki.org/swivt/1.0#Subject>
EQ	<http://www.rdfabout.com/rdf/schema/usgovt/State>	<http://rdf.geospecies.org/ont/geospecies#State>
EQ	<http://data.linkedmdb.org/resource/movie/country>	<http://rdf.geospecies.org/ont/geospecies#Country>
	<http://xmlns.com/foaf/0.1/Person>	<http://semantic-mediawiki.org/swivt/1.0#Subject>
	<http://sw.opencyc.org/2008/06/10/concept/Mx4rqEYnNVMqEdaSKAACs0x8nw>	<http://www.w3.org/2002/07/owl#Class>
	<http://semantic-mediawiki.org/swivt/1.0#Subject>	<http://discoursedb.org/wiki/Special:URIResolver/Category-3ASources>
	<http://discoursedb.org/wiki/Special:URIResolver/Category-3ASources>	<http://semantic-mediawiki.org/swivt/1.0#Subject>
ERR	<http://xmlns.com/foaf/0.1/PersonalProfileDocument>	<http://xmlns.com/foaf/0.1/Person>

Table 3. List of 20 class pairs in CLS-SAME dataset

The above observations about the class pairs in the CLS network reflect that the BTC 2010 dataset is quite heterogenous and the current Semantic Web vocabularies are largely orthogonal. They also enlighten the potential use of the CLS network: with effective classification techniques, we may appropriately label class pairs in the CLS network and then support automated class alignment and error detection. In our future work will also try other combinations of assumptions including the assumption that owl:sameAs is transitive.

7. Related Work

Various recent literature [4-6] investigating pragmatic issues of owl:sameAs in the context of the Web of Data can be considered as directly related to our study. They provide valuable insights into the essential research problem of whether the ubiquitous use of owl:sameAs to inter-linked datasets is correct. Some of them identify incorrect usages of owl:sameAs in the Web of Data [5], leading to the explicit need for a co-reference management infrastructure for the Semantic Web. Others carry out in-depth discussions of the issues with the current semantics of owl:sameAs. For example, McCusker and McGuinness [6] discuss how and why using owl:sameAs could possibly result in confusions of provenance and ground truths in the bioinformatics context. Halpin and Hayes [4] view owl:sameAs statements as a special type of "identity link", and analyze the more general problem of identity links on the Semantic Web from a philosophical and knowledge representation perspective. They also outline four alternative interpretations of owl:sameAs, which all differ from the canonical OWL semantics as defined by W3C documents. Our work differs from all of the above in that, to the best of our knowledge, we are the first to conduct this type of large-scale empirical study on the deployment status of owl:sameAs using datasets from the Web of Data.

Another related research effort is the analysis of the graph structure of the Semantic Web. Some recent work [13-17] presents important graph metrics that reflect the basic shape, structure, and even dynamics of the whole Semantic Web viewed as a giant graph. It is reported in [14] that ontologies on the Semantic Web, like many natural and social networks, are scale-free. Some earlier [16, 17] and later [15] studies show more structural features of the Semantic Web, such as size, diameter and power-law degree distribution of the graph. In one of the more recent efforts that falls into this category, Ge *et al* [13] propose the notion of an Object Link Graph (OLG) for the Semantic Web, and show that it is also scale-free and has a small diameter. Our work is similar to these research efforts in the sense that we also present critical graph structure metrics. However, the subject of research focus, i.e., the owl:sameAs statements, and the scale are two major factors that differentiate our work with theirs.

Some of the existing endeavors, which make use of instance-level links to derive potential alignments and associations at the schema level, are also related to our work. Qu et al [18] propose the notion of a Class Association Graph (CAG), which is obtained from the Object Link Graph (OLG) defined in [13]. Similarly, Nikolov et al [19] illustrate how to establish schema-level mappings based on existing instance-level mappings in the Web of Data. Our study shares essentially the same idea of deriving schema-level relations using vast amounts of instance-level data.

8. Conclusion and Future Work

In order to better understand and use owl:sameAs in Linked Data, it is useful to study how owl:sameAs is actually deployed, which has implications for how data should be consumed. To the best of our knowledge, this work is the first study on SameAs

networks extracted from the real world Web of Data, and it has reported statistically significant results based on the BTC 2010 dataset. The experiment results are the core of this work, and they support the goal of this paper – to highlight the uniqueness, interestingness and utility of SameAs networks to Linked Data researchers as well as practitioners.

- Section 4 shows that SameAs networks have unique graph properties in comparison with other networks in the Semantic Web. The graph properties also lead to nice computational properties of the SameAs network.
- Section 5 explains the interestingness of SameAs networks by showing the similarity between the PLD network and the LOD graph. We also showed that the PLD network could be used to explain how LOD datasets are actually linked by common topics.
- Section 6 shows one practical use of SameAs networks, where classes can be linked by means of common instances (derived by owl:sameAs inference). The CLS network has a great potential in detecting schema-level inconsistencies in interlinked datasets and supporting ontology alignment.

The results reported in this study can be easily extended with additional data, semantics and applications. For example, we can enrich the ESameNet dataset with SameAs statements generated using OWL inference on the entire BTC dataset (e.g. inferring owl:sameAs using owl:InverseFunctionalProperty) [11] and then evaluate the impact on the diameter of SameAs networks. Although this study does not assume the transitivity of owl:sameAs for the purpose of deriving the CLS network, future work may explore the alternative - evaluating the impact of transitive inference on SameAs networks. Another potential research direction is to follow up on our previous discussions on the operational semantics of owl:sameAs [12]. Last but not least, it is worth noting that owl:sameAs has implications not only for the two networks mentioned in this study, but rather, we can use BTC datasets from consecutive years to evaluate the evolution of SameAs Networks over time, and use owl:sameAs statements to compute property-level mappings.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 5(3), Pages 1-22, 2009.

[2] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference. W3C Recommendation, February 2004.

[3] R. Cyganiak. Linked data at the New York Times: Exciting, but buggy. http://dowhatimean.net/2009/10/linked-data-at-the-new-york-times-exciting-but-buggy. Last retrieved September 2010.

[4] H. Halpin and P. J. Hayes. When owl:sameAs isn't the same: An analysis of identity links on the semantic web. In Proceedings of the International Workshop on Linked Data on the Web, 2010.

[5] A. Jaffri, H. Glaser, and I. Millard. URI disambiguation in the context of linked data. In Proceedings of the 1st International Workshop on Linked Data on the Web, 2008.

[6] J. McCusker and D. L. McGuinness. owl:sameAs considered harmful to provenance. In Proceedings of the ISCB Conference on Semantics in Healthcare and Life Sciences, 2010.

[7] B. Vatant. Using owl:sameas in linked data. http://blog.hubjects.com/2007/07/using-owlsameas-in-linked-data.html. Last retrieved September 2010.
[8] H. Lee, D. Leonard, X. Wang, and , D. Loguinov. IRLbot: scaling to 6 billion pages and beyond. In Proceeding of the 17th international Conference on World Wide Web, 2008.
[9] L. Ding, L. Zhou, T. Finin, and A. Joshi, How the Semantic Web is Being Used:An Analysis of FOAF Documents, HICSS38, 2005
[10] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11): 2498–504. 2003.
[11] G. T. Williams, J. Weaver, M. Atre, and J. A. Hendler. Scalable Reduction of Large Datasets to Interesting Subsets. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 8, 2010.
[12] L. Ding, J. Shinavier, T. Finin and D. McGuinness. owl:sameAs and Linked Data: An Empirical Study. Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, 2010.
[13] W. Ge, J. Chen, W. Hu and Y. Qu. Object Link Structure in the Semantic Web. In Proceedings of the 7th Extended Semantic Web Conference (ESWC), 2010.
[14] H. Zhang. The Scale-Free Nature of Semantic Web Ontology. In Proceeding of the 17th international conference on World Wide Web (WWW), 2008.
[15] Y. Theoharis, Y. Tzitzikas, D. Kotzinos, and V. Christophides. On Graph Features of Semantic Web Schemas. IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 5, May 2008.
[16] L. Ding, and T. Finin. Characterizing the Semantic Web on the Web. In Proceedings of the 5th International Semantic Web Conference, 2006.
[17] L. Ding. Enhancing Semantic Web Data Access. Ph.D Thesis. Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, 2006.
[18] Y. Qu, W. Ge, G. Cheng, and Z. Gao. Class Association Structure Derived From Linked Objects. In Proceedings of the WebSci'09: Society On-Line, 2009.
[19] A. Nikolov, V. Uren, and E. Motta. Data Linking: Capturing and Utilising Implicit Schema-level Relations. In Proceedings of the Linked Data on the Web Workshop, 19th International World Wide Web Conference (WWW), 2010.