Assessing trust with PageRank in the Web of Data

José-M. Giménez-Garcia, Harsh Thakkar, Antoine Zimmermann

Assessing Trust with PageRank in the Web of Data

José M. Giménez-García[1], Harsh Thakkar[2], Antoine Zimmermann[3]

[1] Univ Lyon, UJM-Saint-Étienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Étienne, France
jose.gimenez.garcia@univ-st-etienne.fr
[2] Enterprise Information Systems Lab, University of Bonn, Germany
hthakkar@uni-bonn.de
[3] Univ Lyon, MINES Saint-Étienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Étienne, France
antoine.zimmermann@emse.fr

Abstract. While a number of quality metrics have been successfully proposed for datasets in the Web of Data, there is a lack of trust metrics that can be computed for any given dataset. We argue that reuse of data can be seen as an act of trust. In the Semantic Web environment, datasets regularly include terms from other sources, and each of these connections express a degree of trust on that source. However, determining *what* is a dataset in this context is not straightforward. We study the concepts of dataset and dataset link, to finally use the concept of Pay-Level Domain to differentiate datasets, and consider usage of external terms as connections among them. Using these connections we compute the PageRank value for each dataset, and examine the influence of ignoring predicates for computation. This process has been performed for more than 300 datasets, extracted from the LOD Laundromat. The results show that reuse of a dataset is not correlated with its size, and provide some insight on the limitations of the approach and ways to improve its efficacy.

Keywords: linked data, trust, reuse, interlinking, PageRank, metric, assessment

1 Introduction

The WDAqua project[1] aims at advancing the state of the art in data-driven question answering, with a special focus on the Web of Data. The Web of Data comprises thousands of datasets about varied topics, interrelated among them, which contain large quantities of relevant data to answer a question. Nonetheless, in an environment of information published independently by many different actors, data veracity is usually uncertain [17, 19], and there is always the risk of consuming misleading data. While some quality metrics have been proposed

[1] http://wdaqua.informatik.uni-bonn.de/

that can help to identify good datasets [5], there is a lack of trust metrics to provide a confidence on the veracity of the data [23].

In this context, we argue that actual usage of data can be seen as an act of trust. In this paper we focus on reuse of resources by other datasets as a usage metric. We consider reuse of a resource of a dataset by any other given dataset as an outlink from the later to the former. Under this purview, we can compute the PageRank [18] value of each dataset and rank them according to their reuse. PageRank has been successfully used to obtain trust metrics on individual triples [2]. In order to obtain a good measure of reuse, we perform the process on a large scale. We make use of the tools provided by the LOD Laundromat [20] to go beyond LOD Cloud, and process more than 38 billion triples, distributed in more than 600 thousand documents. The LOD Laundromat provides data from data dumps collected from the Internet, so it is not limited to dereferenceable linked data. However, what is regarded as a dataset is an important issue when dealing with data dumps. We make use of the concept of Pay-Level Domain (or PLD, also known as Top-Private Domain) to draw a distinction between datasets, and consider the influence of ignoring predicates when extracting outlinks. We perform a grouping of the triples in datasets according to their PLD and compute their PageRank values as a first measure of trust. Finally, we discuss the results and limitations of the approach, suggesting improvements for future work.

This document is organized as follows: in Section 2, we first discuss the relation of trust and popularity in the Web of Data, what should be considered a dataset in our context in order to clarify the problem we address, and finally present the LOD Laundromat; Section 3 describes the experiments and results, which we further discuss; Section 4 presents relevant related work; finally, we provide some conclusions and directions for future work in Section 5.

2 Ranking the Web of Data

2.1 PageRank, Reuse, and Trust in the Web of Data

We would like to assess trust in datasets by measuring their popularity based on the reuse of resources from a dataset in another dataset. To do this, we rely on the PageRank algorithm [18]. PageRank is the original algorithm developed by Page et al. that Google uses to rank their search results. It takes advantage of the graph structure of the web, considering each link from one page as a "vote" from the source to the destination. Using the links, the importance of a page is propagated across the graph, dividing the value of a page among its outlinks. This process is repeated until convergence is reached. The final result of PageRank corresponds to a stationary distribution, where each page value amounts to the probability for a random surfer to be at any moment in the page.

PageRank is meant to measure popularity (*i.e.*, "human interest and attention") on web pages. However, we argue that reuse of resources in the Semantic

Web has a slightly different meaning. When there is a link from one web page to another, it does not mean necessarily that the author considers the linked page a trustworthy fact (it could be even linking something the author is criticizing). However, resources are reused to express facts in the author's dataset, which implicitly means that the author trusts that the resource is correct. This is supported by the analysis of predicates used for linking datasets by Schmachtenberg et al. [22], where the top used predicates are used to express statements about identity or relatedness (`owl:sameAs`, `rdfs:seeAlso`, `skos:exactMatch`, `skos:closeMatch`), authorship (using Doublin Core vocabulary), and social relations (using `foaf` and `sioc` vocabularies).

To compute PageRank in a set of datasets, it is first necessary to define what is considered a dataset and what is a link between datasets. RDF graphs, although formally defined as a set of triples, can be seen as directed multigraphs in which predicates play the role of arcs. This view suggests that if a triple contains a resource of dataset A as subject, and a resource of dataset B as object, it can be seen as a link from dataset A to dataset B. However, the links formed by arcs in an RDF graph are irrelevant to the notion of dataset linking. In fact, only the presence of hyperlinks suffices to indicate a link between one source and destination, therefore any HTTP IRI in an RDF graph can be seen as a link. So the question is, what it means that a resource belongs to a dataset, and to what dataset a hyperlink "points to". A naïve approach would be to consider that any IRI existing in a dataset belongs to the dataset and thus, that links connect two datasets having one same resource. However, this would imply, for instance, that any triple anywhere that uses a DBpedia IRI is considered to be linked to from the DBpedia dataset. As a result, any dataset that reuses a DBpedia IRI would increase their PageRank according to this definition.

Alternatively, we could take advantage of the linked data principles which stipulate that IRIs should be addresses pointing to a location on the Web. Again, one could naïvely assume that the location that the address points to is what defines the dataset, that is, the document retrieved when one gets the resource using the HTTP protocol. However, this would lead us, for instance, to define each DBpedia article as an individual dataset.

A second possibility would be to use the domain part of the URL, so datasets are grouped by the same publisher. This approach is taken by Ding and Finin [6] to characterize data in the Semantic Web. This way, it would be easy to determine what dataset is being linked to. Such approach would work well if all datasets were accessible from dereferenceable IRIs. However, there are large portions of the Web of Data that provide access to data dumps only [9, 16]. In this case, the domain of the dump does not necessarily match the domain of the individual IRIs found in the dataset. As an example, the DBpedia dumps are found at `http://downloads.dbpedia.org/` while all DBpedia IRIs start with `http://dbpedia.org/`.

The last approach, is to use the concept of *PLD*, *i.e.*, the subdomain component of a URL followed by a public suffix, to identify a dataset. Then, datasets are grouped not necessarily by the same publisher, but by the same publisher

authority. This approach has already been used by other works [15, 22]. As an example, if a file found at `http://download.dbpedia.org/` contains the following triple:

```
<http://dbpedia.org/wiki/Europe>
        <http://www.w3.org/2002/07/owl#sameAs>
                <http://sws.geonames.org/6255148/>
```

we consider that the dataset having the PLD `dbpedia.org` is linking to the dataset with PLD `geonames.org`. It is important to notice that the source of the link (`dbpedia.org`) is obtained from the URL of the document that contains the triple (`http://download.dbpedia.org/`), not from the subject of the example. This approach enables us to extract outlinks from datasets published in dumps, and therefore access the majority of accessible semantic web data.

Definition 1 (Dataset). *A* dataset *is a non empty collection of triples that can be retrieved from sources accessible at a URL having a common Pay-Level Domain. The PLD identifies the dataset.*

In the previous example, we see that the predicate IRI is linking to the standard OWL vocabulary. It is very likely that predicates in general will be linking to vocabularies that are extensively reused. However, our intent is to evaluate trust on actual data that can be used to answer questions, and not vocabularies used to describe the data. We predict that extracting outlinks from predicates will lead to higher values for datasets containing only vocabularies. For this reason, we perform the same experiment with and without taking predicates into consideration.

Definition 2 (Dataset link). *There exists a* link *from a dataset A to a dataset B if and only if there exists a triple in a file at a location having the PLD that identifies A in which the PLD of its subject, its object, or both matches the PLD that identifies B.*

This definition is in line with the PageRank algorithm [18] where the number of links between the same two nodes is irrelevant. Note that since datasets must be non empty, links to PLDs that do not host RDF have to be ignored.

2.2 The LOD Laundromat and Frank

The LOD cloud[2], and in general Linked Open Data, contains a wide variety of formats, publishing schemes, errors, that make it difficult to perform a large-scale evaluation. Yet, to be accurate, our study requires to be comprehensive. Fortunately, the LOD Laundromat [1, 21] makes this data available by gathering dataset dumps from the Web, including archived data. LOD Laundromat cleans the data by fixing syntactic errors and removing duplicates, and then makes it available through download (either as gzipped N-Triples or N-Quads,

[2] `http://lod-cloud.net/`

or HDT [10] files), a SPARQL endpoint, and Triple Pattern Fragments [24]. Using the LOD Laundromat is also a better solution than trying to use documents dereferenced by URIs, because most of datasets available online are data dumps [9, 16], thus not accessible by dereferencing.

Frank [20] is a command-line tool which serves as an interface of the LOD Laundromat, and makes it easy to run evaluations against very large numbers of datasets.

3 Experiments and Results

The process to compute PageRank involves the following steps, detailed further below and illustrated in Figure 1. The code and results are provided online.[3]

1. Extracting the document list from LOD Laundromat.
2. Parsing the content of each document to extract the outlinks.
3. Consolidating the results
4. Computing PageRank

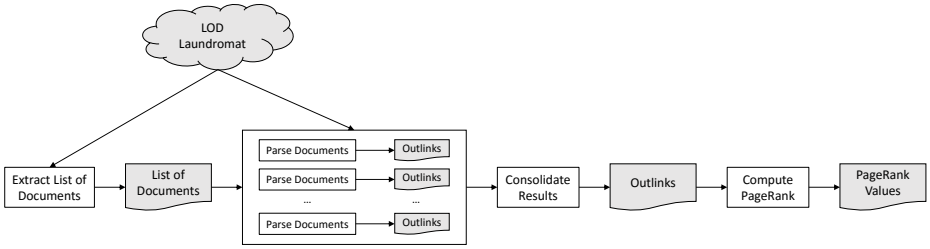

Fig. 1: Outlink extraction and PageRank computation workflow

3.1 Extracting the document list from LOD Laundromat

We use the Frank command line tool [20] to obtain a snapshot of the contents of the LOD Laundromat. While the output of Frank can be directly pipelined to our process, the next step is performed in parallel in several machines. For this reason, we need that every machine reads the exact same input. An update in the contents of the LOD Laundromat during the next process could have impacted the results in that case. We retrieve the list of documents in the LOD Laundromat with the following command.

```
$ frank documents > documents.dat
```

This command retrieves a list of pairs *(downloadURL-resourceURL)*, where the first is the URL to download the gzipped datasets, and the second the resource identifier in the LOD Laundromat ontology. At the moment of the experiments, it retrieved 649,855 documents.

[3] `https://github.com/jm-gimenez-garcia/LODRank`

3.2 Parsing the content of each document to extract the outlinks.

A prototype tool[4] has been developed to stream the contents of the documents end extract the outlinks. This tool reads the list of pairs *(downloadURL-resourceURL)* from the standard input, and accepts two optional parameters for partial processing: *Step* and *Start*. The first one tells how many lines the process reads in every iteration, processing the last one, while the second denotes what line to use for the first input. For each line processed, it queries the SPARQL endpoint to retrieve the URL where that datasets was crawled. This information can be found in the LOD Laundromat ontology connected to the resource, in the case the document was crawled as a single file, or connected to the archive that contains the document, if it was crawled compressed in a compressed file, possibly along other documents. In the first case, we retrieve the URL with Query 1, in the second case we retrieve the URL using Query 2, where `%s` is substituted by the *resourceURL*. The Pay-Level Domain is then extracted and stored. This will be considered as the identifier of the dataset.

```
SELECT ?url
WHERE {<%s> <http://lodlaundromat.org/ontology/url> ?url}
```

Query 1: Query to retrieve crawled URL of a non-archived document

```
SELECT ?url
WHERE {
       ?archive <http://lodlaundromat.org/ontology/containsEntry> <%s> .
       ?archive <http://lodlaundromat.org/ontology/url> ?url
}
```

Query 2: Query to retrieve crawled URL of an archived document

Then, the gzipped file is streamed from the *downloadURL* and parsed the triples. The subject and object (in case it is a URI) are extracted the Pay-Level Domain and compared against their dataset PLD. If they have a valid PLD and is different from their dataset's Pay-Level Domain, the pair *(datasetPLD-resourcePLD)* is stored as an outlink for the dataset. The output of each dataset is stored in a different file, which will be appended more pairs if a different document is identified as the same dataset.

[4] `https://github.com/jm-gimenez-garcia/LODRank/tree/master/src/com/chemi2g/lodrank/outlink_extractor`

This process makes use of Apache Jena[5] v3.0.1 to query the SPARQL endpoint of the LOD Laundromat and Google Guava[6] v19.0 to extract the Pay-Level Domain of the datasets.

In the experiments the process was launched in parallel in 8 virtual machines using Google Cloud Platform[7] free trial resources, each one processing a different subset of the list downloaded in the previous step. A statistical description of the results of each process, with and without considering predicates, is detailed in Table 1. "Documents" correspond to the number of dump files in the LOD Laundromat, while "Datasets" are the number of PLDs that the process is dealing with. There can be an overlap in the datasets of several processes, so the total number of datasets is not equal to the sum. We can see that the number of triples processed by each process is not proportional to the number of documents processed.

Process	Documents	Triples	Datasets (w. p.)	Datasets (w/o. p.)
1	81,220	3,994,446,393	135	121
2	81,226	3,742,870,561	137	118
3	83,422	4,146,249,367	140	127
4	81,225	3,376,784,600	135	120
5	81,225	3,623,413,245	142	120
6	88,198	3,377,773,585	131	116
7	81,226	4,132,960,522	137	115
8	89,781	3,911,917,919	134	123

Table 1: Data extracted from the LOD Laundromat by each process

3.3 Consolidating the results

Once the outlinks have been extracted, the different files have to be appended and duplicates removed using a simple tool.[8] In the experiments, the data from each virtual machine was downloaded in a separate folder of a unique machine. Then files with the same name in each folder were concatenated and we removed the duplicates. The total number of datasets after consolidating the results is 412 when considering predicates, and 319 when not. The result was again concatenated in a single file.

[5] `https://jena.apache.org/`
[6] `https://github.com/google/guava`
[7] `https://cloud.google.com/`
[8] `https://github.com/jm-gimenez-garcia/LODRank/tree/master/src/com/chemi2g/lodrank/duplicate_remover`

3.4 Computing PageRank

For PageRank computation we make use of the igraph R package [4]. The ordered PageRank values for all datasets can be seen in Figure 2 and Figure 3, with a logarithmic scale. The complete list of results is published online.[9] We can see that in both cases the top-ranked dataset is very much higher than the rest, then the slope becomes more regular until it reaches a plateau at the end, with a minimum value shared by several datasets that have no inlinks at all. Tables 2 and 3 show the 10 highest ranked datasets.

Discussion. Here we provide additional information about the datasets, especially the top-ranked ones, in order to understand how ranking correlates with other statistical values, such as number of triples, number of documents. We also discuss how our own choices influenced the results.

The datasets appearing on the top 10 list are generally not surprising, with the only exception of `holygoat.co.uk`, the only domain in the top 10 owned by an individual person, Richard Newman, a computer scientist who wrote several ontologies in the early days of the Semantic Web. This is even more remarkable considering that the dataset has only 7 inlinks. The reason is that `rdfs.org`, the domain of `sioc` ontology for instance, includes resources from `holygoat.co.uk`. Because this dataset has only 2 outlinks, half of its PakeRank score is forwarded to `holygoat.co.uk`, which accounts for 96% of its PageRank value.

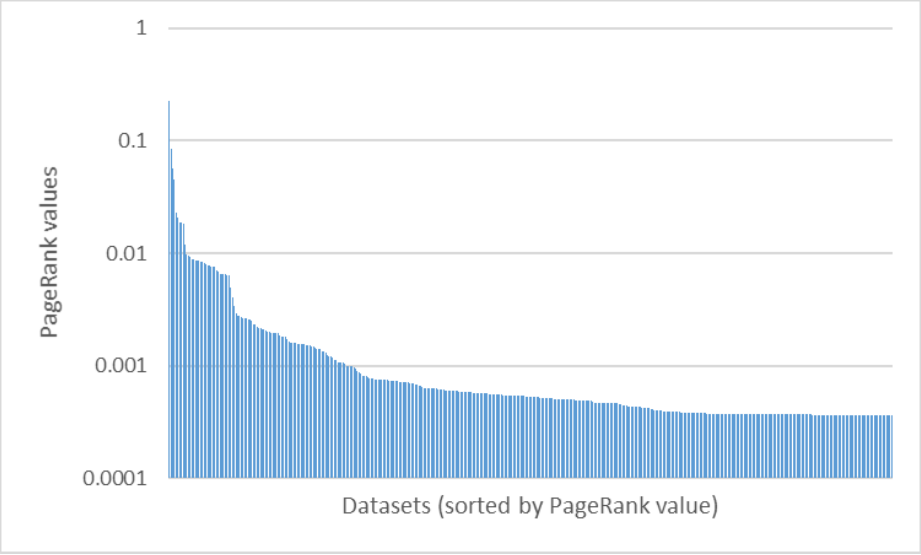

Fig. 2: PageRank values sorted from higher to lower, with predicates

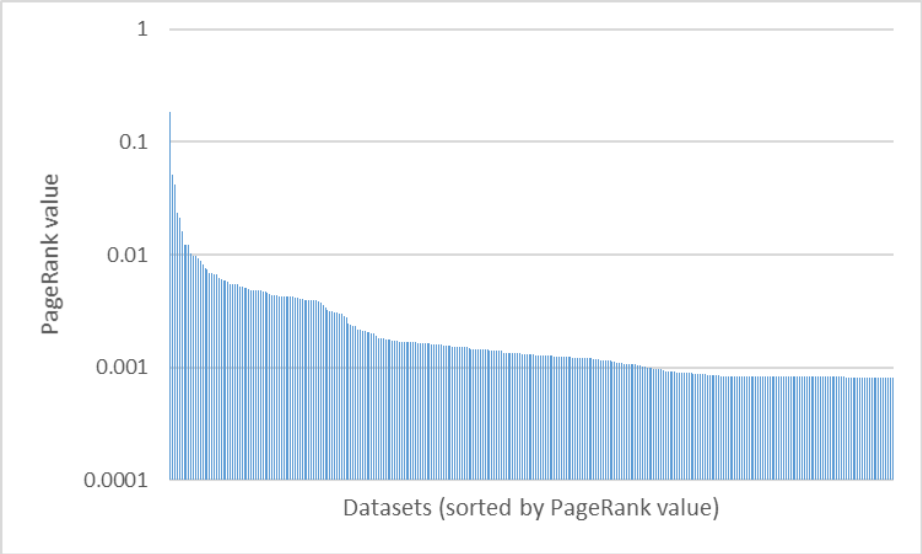

Fig. 3: PageRank values sorted from higher to lower, without predicates

Rank	Dataset	PageRank value	Inlinks	Outlinks
1	w3.org	0.224806691	411	32
2	purl.org	0.085548846	278	64
3	lodlaundromat.org	0.056963209	188	1
4	xmlns.com	0.045452453	219	3
5	schema.org	0.023239532	32	1
6	creativecommons.org	0.020496922	106	2
7	dbpedia.org	0.018894825	118	160
8	rdfs.org	0.018738995	108	5
9	ogp.me	0.018442606	37	4
10	usefulinc.com	0.012066847	26	4

Table 2: PageRank values for the top 10 datasets, with predicates

Rank	Dataset	PageRank value	Inlinks	Outlinks
1	purl.org	0.185304616	181	50
2	creativecommons.org	0.051625742	93	1
3	dbpedia.org	0.04234706	104	119
4	rdfs.org	0.023497322	73	2
5	geonames.org	0.02127494	59	6
6	loc.gov	0.016137225	33	8
7	fao.org	0.012392539	27	8
8	europa.eu	0.012182709	30	13
9	holygoat.co.uk	0.012179038	7	1
10	data.gov.uk	0.010364034	19	11

Table 3: PageRank values for the top 10 datasets, without predicates

As predicted, when including predicates the first positions incorporate more datasets about vocabularies. When removing the predicates, `w3.org`, `xmlns.com`, `schema.org`, and `ogp.me` no longer appear in the top positions, and datasets with factual data move upwards. `lodlaundromat.org` seems to appear when considering predicates because the LOD Laundromat adds information about the cleaning process when processing the data. While not an optimum solution (considering that `purl.org` and `rdfs.org`, hosts of well known ontologies, are still in the top positions), ignoring the predicates proves to be a simple but useful technique.

We used two queries, (Query 3 and Query 4), to obtain the number of documents and triples for each PLD, from the LOD Laundromat.

```
PREFIX llo: <http://lodlaundromat.org/ontology/>
PREFIX ll: <http://lodlaundromat.org/resource/>
SELECT (COUNT(DISTINCT ?resource) AS ?count)
WHERE {
       {
              ?resource llo:url ?url
              FILTER regex(?url, "[^/\\.]*\\.?%s/", "")
       }
       UNION
       {
              ?archive llo:containsEntry ?resource ;
                     llo:url ?url
              FILTER regex(?url, "[^/\\.]*\\.?%s/", "")
       }
}
```

Query 3: Query to retrieve the number of documents per dataset

The result of the queries are given in Table 4 for all the datasets that appear in the 10 top of both experiments.

As we can see, popularity is not at all correlated with the size of the datasets. Indeed, a number of the top ten datasets have less that 200 triples, while `dbpedia.org` and `europa.eu` both have billions of triples.

The enormously high page rank of `purl.org` should be mitigated by the fact that `purl.org` does not actually host any data. It is a redirecting service that many data publishers are using. This result highlights a drawback in our heuristic for identifying datasets: the PLD is not always referring to a single dataset. To overcome this particular case, we could consider the PLD of the URL of the document obtained after dereferencing the IRI, in the same way as Hogan et al. [15] do for the general case.

[9] `https://github.com/jm-gimenez-garcia/LODRank/tree/master/results`

```
PREFIX llo: <http://lodlaundromat.org/ontology/>
PREFIX ll: <http://lodlaundromat.org/resource/>
SELECT (COUNT(DISTINCT ?resource) AS ?count) (SUM(?triples) as ?sum)
WHERE {
       {
              ?resource llo:url ?url ;
                     llo:triples ?triples
       FILTER (?triples > 0)
              FILTER regex(?url, "[^/\\.]*\\.?%s/", "")
       }
       UNION
       {
              ?archive llo:containsEntry ?resource ;
                     llo:url ?url .
              ?resource llo:triples ?triples
       FILTER (?triples > 0)
              FILTER regex(?url, "[^/\\.]*\\.?%s/", "")
       }
}
```

Query 4: Query to retrieve the number of documents with triples and number of triples

Dataset	Rank	Documents	Documents with triples	Triples
w3.org	1 / -	413	256	1,973,715
purl.org	2 / 1	9,166	9,073	254,548,441
lodlaundromat.org	3 / -	68	1	4
xmlns.com	4 / -	4	4	1895
schema.org	5 / -	1	1	1
creativecommons.org	6 / 2	1	1	117
dbpedia.org	7 / 3	1,888	1,752	1,257,930,891
rdfs.org	8 / 4	6	6	1,808
ogp.me	9 / 274	68	1	231
usefulinc.com	10 / -	2	2	1398
geonames.org	20 / 5	6	4	9,762
loc.gov	29 / 6	16	12	263,653,979
fao.org	36 / 7	17	11	48366
europa.eu	12 / 8	7734	7,705	3,414,066,228
holygoat.co.uk	37 / 9	1	1	95
data.gov.uk	42 / 10	157	88	51,401,490

Table 4: Documents and triples per dataset in LOD Laundromat

Another possible drawback of the approach is that triples with `rdf:type` in predicate position have their object pointing to a class in an ontology. This is in contradiction with our remark in Section 2 where we say that we want to rank instance data rather than terminological knowledge. This can have a major impact the results since `purl.org` is most often used to redirect to vocabularies more than datasets, and `rdfs.org` only hosts ontologies.

4 Related work

The authors of Semantic Web Search Engine (SWSE [15]) strongly advocate that the use of a ranking mechanism is very crucial for prioritizing data elements in the search process. Their work is inspired by the Google PageRank algorithm, which treats hyperlinks to other pages as a positive score. The PageRank algorithm is targeted for hyperlink documents and its adaptation to the LOD is however non-trivial, as we have seen. They point out that the primary reason for this is that LOD datasets may not have direct hyperlinks to other datasets but rather in most cases make use of implicit links to other web pages via the re-use of dereferenceable URIs. Here the unit of search becomes the entity and not the document itself. The authors briefly re-introduce the concept of naming authority, from their previous work [13] in order to rank structured data from an open distributed environment. They assume that the naming authority should match the Pay-level domain such that computing PageRank is performed on a naming authority graph where the nodes are PLDs. Their intuition therefore is in accordance with our reasoning from Section 2. They have discussed and contrasted the interpretation of naming authorities on a document level (e.g. `http://www.danbri.org/foaf.rdf`) and a PLD level (`danbri.org`). Also, they make use of a generalization for the method discussed in the paper [8] for ranking entities and carry out links analysis on the PLD abstraction layer.

The authors of Swoogle [7] develop OntoRank algorithm in order to rank documents. OntoRank, a variation of Google PageRank, is an iterative algorithm for calculating the ranks for documents built on references to terms (i.e., classes and properties) which are defined in other documents.

In the paper [3], the authors calculate the rank of entities (or as they call them objects) based on the logarithm of the number of documents where that particular object is mentioned.

In their work [11] present LinkQA, an extensible data quality assessment framework for assessing the quality of linked data mappings using the network measures. For this, they assess the degree of interlinking of datasets using five network measures, out of which two network measures are specifically designed for Linked Data (namely, Open Same-As chains and Description Richness) and the other three standard network measures (namely, degree, centrality, and the clustering coefficient) in order to assess variation in the quality of the overall linked data with respect to a certain set of links.

In [2], PageRank is used to compute a measure that is in turn associated to individual statements in datasets for the purpose of incorporating trust in

reasoning. Therefore, as in our own approach, they consider that PageRank is an indication of trustworthiness. However, they only compute PageRank on a per document basis, and report on the PageRank values of the top 10 documents obtained from their web crawl.

5 Conclusion & Future work

Data-driven question answering, the aim of project WDAqua mentioned in the introduction of this paper, requires quality data in which one can trust. Our aim has been to provide insight on how a trust measure can be based on dataset interlinking. To that end, we consider Pay-Level Domains as identifiers of unique datasets and compute PageRank on them. Our results show that the design choices greatly affect the results. Whether taking into account or not predicates for outlink extraction impacts how vocabularies are ranked, and the choice of PLD as definition of dataset is arguable, as some PLDs group many data dumps. In order to improve this, we could associate well known datasets to IRI patterns, such as `it.dbpedia.org` for the Italian version of DBpedia.

In addition, we also intend to explore further applications of PageRank that may be useful for question answering. User interaction that provides trust values in a number of dataset could be used to compute PageRank values with those datasets as a teleport set, as suggested by Gyöngyi et al. [12]. Also, Topic-Sensitive PageRank [14] could help a question-answering system to select different datasets when a question is identified to belong to a specific topic.

Finally, this work is part of a broader objective that we want to pursue: to ascertain the relationship between the perceived trust on a dataset and its objective quality. We will explore this area in a future work where other data reuse metrics will be considered and compared against different quality metrics.

Acknowledgement

This project is supported by funding received from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 642795. We would like to thank Elena Simperl, whose idea jumpstarted the project that lead to this article, and also Elena Demidova, Kemele Endris, and Christoph Lange for the useful discussions related to it.

References

[1] Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data. In: The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference. Lecture Notes in Computer Science, vol. 8796, pp. 213–228. Springer (2014)

[2] Bonatti, P.A., Hogan, A., Polleres, A., Sauro, L.: Robust and scalable Linked Data reasoning incorporating provenance and trust annotations. Web Semantics: Science, Services and Agents on the World Wide Web 9(2), 165–201 (2011)

[3] Cheng, G., Qu, Y.: Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. International Journal on Semantic Web and Information Systems (IJSWIS) 5(3), 49–70 (2009)

[4] Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal, Complex Systems 1695(5), 1–9 (2006)

[5] Debattista, J., Londoño, S., Lange, C., Auer, S.: Quality Assessment of Linked Datasets Using Probabilistic Approximation. In: The Semantic Web. Latest Advances and New Domains - ESWC 2015 - 12th Extended Semantic Web Conference. Lecture Notes in Computer Science, vol. 9088, pp. 221–236. Springer (2015)

[6] Ding, L., Finin, T.: Characterizing the Semantic Web on the Web. In: The Semantic Web - ISWC 2006 - 5th International Semantic Web Conference. Lecture Notes in Computer Science, vol. 4273, pp. 242–257. Springer (2006)

[7] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. pp. 652–659. ACM (2004)

[8] Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P.: Finding and Ranking Knowledge on the Semantic Web. In: The Semantic Web - ISWC 2005 - 4th International Semantic Web Conference. Lecture Notes in Computer Science, vol. 3729, pp. 156–170. Springer (2005)

[9] Ermilov, I., Martin, M., Lehmann, J., Auer, S.: Linked Open Data Statistics: Collection and Exploitation. In: Knowledge Engineering and the Semantic Web - KESW 2013 - 4th International Conference. Communications in Computer and Information Science, vol. 394, pp. 242–249. Springer (2013)

[10] Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). Web Semantics: Science, Services and Agents on the World Wide Web 19, 22–41 (2013)

[11] Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Assessing Linked Data Mappings Using Network Measures. In: The Semantic Web: Research and Applications - ESWC 2012 - 9th Extended Semantic Web Conference. Lecture Notes in Computer Science, vol. 7295, pp. 87–102. Springer (2012)

[12] Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating Web Spam with TrustRank. In: Proceedings of the Thirtieth international conference on Very large data bases. vol. 30, pp. 576–587. VLDB Endowment (2004)

[13] Harth, A., Kinsella, S., Decker, S.: Using Naming Authority to Rank Data and Ontologies for Web Search. In: The Semantic Web - ISWC 2009 - 8th International Semantic Web Conference. Lecture Notes in Computer Science, vol. 5823, pp. 277–292. Springer (2009)

[14] Haveliwala, T.H.: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE transactions on knowledge and data engineering 15(4), 784–796 (2003)
[15] Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S.: Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. Web Semantics: Science, Services and Agents on the World Wide Web 9(4), 365–401 (2011)
[16] Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of Linked Data conformance. Web Semantics: Science, Services and Agents on the World Wide Web 14, 14–44 (2012)
[17] Liu, S., d'Aquin, M., Motta, E.: Towards Linked Data Fact Validation through Measuring Consensus. In: Proceedings of the 2nd Workshop on Linked Data Quality co-located with 12th Extended Semantic Web Conference (ESWC 2015). CEUR Workshop Proceedings, vol. 1376. CEUR-WS.org (2015)
[18] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web (1999)
[19] Paulheim, H., Bizer, C.: Improving the Quality of Linked Data Using Statistical Distributions. International Journal on Semantic Web and Information Systems (IJSWIS) 10(2), 63–86 (2014)
[20] Rietveld, L., Beek, W., Schlobach, S.: LOD Lab: Experiments at LOD Scale. In: The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference. Lecture Notes in Computer Science, vol. 9367, pp. 339–355. Springer (2015)
[21] Rietveld, L., Verborgh, R., Beek, W., Vander Sande, M., Schlobach, S.: Linked Data-as-a-Service: The Semantic Web Redeployed. In: The Semantic Web. Latest Advances and New Domains - ESWC 2015 - 12th Extended Semantic Web Conference. Lecture Notes in Computer Science, vol. 9088, pp. 471–487. Springer (2015)
[22] Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data Best Practices in Different Topical Domains. In: The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference. Lecture Notes in Computer Science, vol. 8796, pp. 245–260. Springer (2014)
[23] Thakkar, H., Endris, K.M., Giménez-García, J.M., Debattista, J., Lange, C., Auer, S.: Are Linked Datasets fit for Open-domain Question Answering? A Quality Assessment. In: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016. p. 19. ACM (2016)
[24] Verborgh, R., Vander Sande, M., Colpaert, P., Coppens, S., Mannens, E., Van de Walle, R.: Web-Scale Querying through Linked Data Fragments. In: Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014). CEUR Workshop Proceedings, vol. 1184 (2014)