Assessing Linked Data Mappings using Network Measures

Christophe Guéret[1], Paul Groth[1], Claus Stadler[2], and Jens Lehmann[2]

[1] Free University Amsterdam, De Boelelaan 1105, 1081HV Amsterdam
{c.d.m.gueret,p.t.groth}@vu.nl
[2] University of Leipzig, Johannisgasse 26, 04103 Leipzig
{cstadler,lehmann}@informatik.uni-leipzig.de

Abstract. Linked Data is at its core about the setting of links between resources. Links provide enriched semantics, pointers to extra information and enable the merging of data sets. However, as the amount of Linked Data has grown, there has been the need to automate the creation of links and such automated approaches can create low-quality links or unsuitable network structures. In particular, it is difficult to know whether the links introduced improve or diminish the quality of Linked Data. In this paper, we present LINK-QA, an extensible framework that allows for the assessment of Linked Data mappings using network metrics. We test five metrics using this framework on a set of known good and bad links generated by a common mapping system, and show the behaviour of those metrics.

Keywords: linked data, quality assurance, network analysis

1 Introduction

Linked Data features a distributed publication model that allows for any data publisher to semantically link to other resources on the Web. Because of this open nature, several mechanisms have been introduced to semi-automatically link resources on the Web of Data to improve its connectivity and increase its semantic richness. This partially automated introduction of links begs the question as to which links are improving the *quality* of the Web of Data or are just adding clutter. This notion of quality is particularly important because unlike the regular Web, there is not a human deciding based on context whether a link is useful or not. Instead, automated agents (with currently less capabilities) must be able to make these decisions.

There are a number of possible ways to measure the quality of links. In this work, we explore the use of network measures as one avenue of determining the quality. These statistical techniques provide summaries of the network along different dimensions, for example, by detecting how interlinked a node is within in a network [3]. The application of these measures for use in quality measurement is motivated by recent work applying networks measures to the Web of Data [11].

Concretely, we pose the question of whether network measures can be used to detect changes in quality with the introduction of new links (*i.e.* mappings) between datasets. We test 5 network measures, three classic network measures (degree, centrality, clustering coefficient) and two network measures designed specifically for Linked Data (Open SameAs chains, and Description Richness). The experiments are performed on link sets produced by Silk [23], a state-of-the-art mapping tool. We show that at this time such network measures are only partially able to detect quality links. We discuss reasons for this and sketch a possible path forward.

In addition to these experiments, we present an extensible framework, LINK-QA, for performing such network analysis based quality assessment. The framework allows for both the execution and reporting of quality measurements. Our contributions are twofold:

1. a framework, LINK-QA, for measuring quality of topological modifications to Linked Data; and
2. analysis of five network measures for the applicability in testing link quality.

The rest of this paper is organized as follows. We begin with some preliminary definitions of the the networks we analyze. The metrics we test are defined in Section 3. This is followed 4 by a description of the framework for quality assessment. This includes a discussion of a reference implementation. Experimental results on a set of automatically generated links are reported on in Section 5. Finally, we discuss related work and conclude.

2 Network definitions

We now introduce the definitions used throughout this paper. The graph we want to study will be referred to as the *Data Network*. It is the network of facts provided by the graph of the Web of Data, excluding the blank nodes.

Definition 1 (Data Network). *The data network is defined as a directed, labelled, graph $\mathcal{G} = \{V, E, L\}$ with V a set of nodes, E a set of edges and L a set of labels. An edge e_{ij} connects a node $v_i \in V$ to the node $v_j \in V$ with a label $l(e_{ij}) \in L$. The edges and labels correspond to the triples and predicates of the Web of Data.*

In this paper, we sample the Data Network by collecting information about the neighbourhood of selected sets of resources within it. A resource's neighbourhood consists of a direct neighbourhood and an extended neighbourhood:

Definition 2 (Resource Neighbourhood). *The direct neighbourhood of a node i is defined as the set of nodes directly connected to it through either an incoming edge (N_i^-) or outgoing edge (N_i^+). That is, $N_i = N_i^+ \cup N_i^- = \{v_j \mid e_{ij} \in E\} \cup \{v_j \mid e_{ji} \in E\}$. Its extended neighbourhood N_i^* also include neighbours' neighbours which are not i: $N_i^* = N_i \cup \bigcup_{v_j \in N_i} N_j$.*

A resource neighbourhood is used to build a local network around a particular resource from the Data Network.

Definition 3 (Local Network). *The local network $\mathcal{G}_i = \{V_i, E_i, L_i\}$ of a node $v_i \in V$ is a directed, labeled graph of the extended neighbourhood of v_i. The set of nodes is defined as $V_i = N_i^*$, the edges are $E_i = \{e_{jk} \in E \mid (v_j, v_k) \in N_i^* \times N_i^*\}$ and the labels $L_i = \{l(e_{jk}) \mid e_{jk} \in E_i\}$.*

Figure 1 shows an example of a local network for which ⬤ is v_i, ⬤ are the nodes in N_i and ⬤ are the nodes in N_i^*. Local networks created around nodes from $\mathcal{G}$ are the focus of the analysis performed by our framework. It is worth noting that the union of all local neighbourhoods of every node in $\mathcal{G}$ is equivalent to this graph. That is, $\mathcal{G} \equiv \bigcup_{v_i \in N} \mathcal{G}_i$.
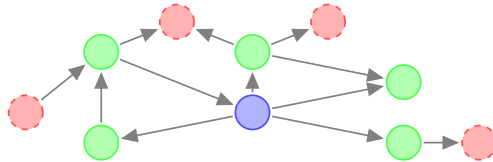

Fig. 1. Example of direct and extended neighbourhood around a source node

3   Network metrics

Based on the above definitions, we now detail a set of 5 network metrics to use in quality detection. Relevant prior work on network analysis was the key criteria to establish these metrics. Although Linked Data networks are different to social networks, we used them as starting point. The degree, clustering coefficient and centrality measures are justified as measures of network robustness [1,10]. The other two metrics are based on studies of the Web of Data as a network that show that fragmentation of the SameAs network is common and thus may be a sign of low quality [12].

In specifying these metrics, one must not only define the measure itself but also what constitutes quality with respect to that measure. Defining such a "quality goal" is difficult as we are only beginning to obtain empirical evidence about what network topologies map to qualitative notions of quality [10]. To address this problem, for each metric, we define an ideal and justify it with respect to some well-known quality notions from both network science and Linked Data publication practice. We consider the following to be broad quality goals that should be reached by the creation of links: 1. modifications should bring the topology of the network closer to that of a power law network to make the network more robust against random failure; 2. modifications should lower the differences between the centrality of the hubs in the network to make the network more robust against targeted failure of these critical nodes; 3. modifications should increase the clustering within topical groups of resources and also lower the average path length between groups (*i.e.* foster a small world network).

We now discuss the 5 metrics (degree, clustering coefficient, open sameAs chains, centrality, description richness). We describe the measure itself as well as the ideal (*i.e.* goal) distribution associated to it. Each metric is designed to fulfill the following criteria:

1. be computable using the local network of a resource;
2. be representative of a global network property;
3. be able to identify particular parts of the network that relate to an ideal distribution of that metric;
4. have a domain within positive real values (the metrics described here produce values between 0 and a factor of N).

We note that local networks can have variable size and thus may not be independent of each other as there may be overlaps. The network measures themselves are aggregations of those at the local level. In the development of these metrics, we attempt to ensure that the metrics are not sensitive to local level effects.

3.1 Degree

This measures how many hubs there are in a network. The aim is to have a network which allows for fast connectivity between different parts of the network. Thus making it easier for automated agents to find a variety of information through traversal. Power-law networks are known to be robust against random failure and are a characteristic of small world networks [1].

Measure The degree of a node is given by its number of incoming and outgoing edges.

$$m_i^{degree} = \|\{e_{ij} \mid v_j \in N_i, e_{ij} \in E_i\}\| + \|\{e_{ji} \mid v_j \in N_i, e_{ji} \in E_i\}\|$$

Ideal We aim at a degree distribution that follows some power-law $P(k) \sim ck^{-\gamma}$ where $P(k)$ is the probability of finding a node with a degree k and c, γ are two distribution parameters. Power-law degree distributions are a sign of robustness against random failure and one of the characteristics of small world networks. The distance between the degree distribution and its ideal is defined as the absolute difference between the observed distribution and its closest power-law equivalent obtained through fitting.

$$d^{degree} = \sum_k abs(\frac{\|\{v_i \mid m_i^{degree} = k\}\|}{\|N_i\| + 1} - ck^{-\gamma})$$

3.2 Clustering Coefficient

The clustering coefficient is a measure of the denseness of the network. The metric measures the density of the neighbourhood around a particular node.

Measure The local clustering coefficient of a node n_i is given by the ratio between the number of links among its neighbours and the number of possible links.

$$m_i^{clustering} = \frac{\|\{e_{jk} \mid v_j, v_k \in N_i, e_{jk} \in E_i\}\|}{\|N_i\|(\|N_i\| - 1)}$$

Ideal The highest average clustering coefficient a network can have is 1, meaning that every node is connected to every other node (the network is said to be "complete"). Although this is a result the Web of Data should not aim at, as most links would then be meaningless, an increase of the clustering coefficient is a sign of cohesiveness among local clusters. The emergence of such topic oriented clusters are common in the Web of Data and are in line with having a small world. We thus set an average clustering coefficient of 1 as a goal and define the distance accordingly. S being the set of all resources, the distance to the ideal is 1 minus the average clustering coefficient of the nodes in S.

$$d^{clustering} = 1 - \frac{1}{\|S\|} \sum_{v_i \in S} m_i^{clustering}$$

3.3 Centrality

Commonly used estimates of the centrality of a node in a graph are *betweenness centrality*, *closeness centrality*, and *degree centrality*. All these values indicate the critical position of a node in a topology. For this metric, we focus on betweenness centrality, which indicates the likelihood of a node being on the shortest path between two other nodes. The computation of betweenness centrality requires knowing the complete topology of the studied graph. Because our metrics are node-centric and we only have access to the local neighbourhood, we use the ratio of incoming and outgoing edges as a proxy.

Measure The centrality of a node v_i is given by the number of connections it takes part in. This value is obtained by the product between the number of nodes reaching v_i through its incoming neighbours, and the number of nodes reachable through the outgoing neighbours.

$$m_i^{centrality} = \frac{\|\{v_k \mid e_{kj} \in E_i, v_j \in N_i^+\}\|}{\|\{v_k \mid e_{jk} \in E_i, v_j \in N_i^-\}\|}$$

Ideal A network dominated by highly central points is prone to critical failure in case those central points cease to operate or are being renamed [10]. Ideally, the creation of new links would reduce the overall discrepancy among the centrality values of the nodes. This means decreasing the centrality index of the graph:

$$d^{centrality} = \sum_{i \in V} \frac{max_{j \in V}(m_j^{centrality}) - m_i^{centrality}}{\|V\| - 1}$$

3.4 SameAs Chains

The very common `owl:sameAs` property can be improperly asserted. One way to confirm a given sameAs relation is correct is to find closed chains of sameAs relations between the linking resource and the resource linked. This metric detects whether there are open sameAs chains in the network.

Measure The metric counts the number of sameAs chains that are not closed. Let $p_{ik} = \{e_{ij_1}, \ldots, e_{j_y k}\}$ be a path of length y defined as a sequence of edges with the same label $l(p_{ik})$. The number of open chains is defined as

$$m_i^{paths} = \|\{p_{ik} \mid l(p_{ik}) = \text{"owl:sameAs"}, k \neq i\}\|$$

As highlighted earlier, metrics should not be sensitive on scaling effects when going from the local definition to their global value. This metric is not sensitive under the assumption that there are few long sameAs chains in the global network [12].

Ideal Ideally, we would like to have no open sameAs chains in the WoD. If the new links contribute to closing the open paths, their impact is considered positive.

$$d^{paths} = \sum_{v_i \in V} m_i^{paths}$$

3.5 Descriptive Richness through SameAs

This metric measures how much to the description of a resource is added through the use of sameAs edges. If a sameAs edge is introduced, we can measure whether or not that edge adds to the description.

Measure The measure counts the number of new edges brought to a resource through the sameAs relation(s). This initial set of edges is defined as $A_i = \{e_{ij} \mid l(e_{ij}) \neq \text{"owl:sameAs"}, j \in N_i^+\}$ the number of edges brought to by the neighbours connected through a sameAs relation defined as $B_i = \{e_{jl} \mid v_l \in N_j^+, l \neq i, e_{ij} \in N_i^+, l(e_{ij}) = \text{"owl:sameAs"}\}$ Finally, the gain is the difference between the two sets

$$m_i^{description} = B_i \setminus A_i$$

Ideal A resource's outgoing sameAs relations ideally link to resources that have a complementary description to the original one. Therefore, the richer the resulting description, the lower the distance to our ideal.

$$d^{description} = \sum_{i \in V} \frac{1}{1 + m_i^{description}}$$

4 LINK-QA analysis framework

The above metrics were tested using LINK-QA, a framework for assessing the quality of Linked Data using network metrics. The framework is scalable and extensible: the metrics applied are generic and share a common set of basic requirements, making it easy to incorporate new metrics. Additionally, metrics are computed using only the local network of a resource and are thus parallelisable by design. The framework consists of five components, "**Select**", "**Construct**", "**Extend**", "**Analyse**" and "**Compare**". These components are assembled together in the form of a workflow (see Figure 2).
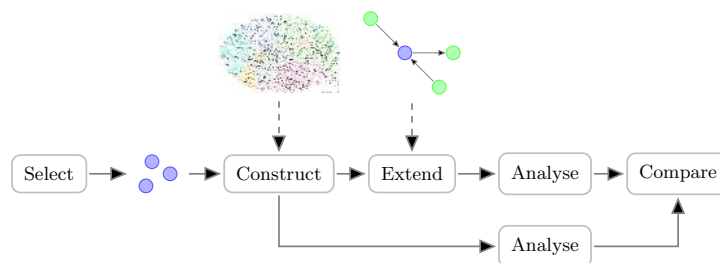

Fig. 2. Interaction between the different components of LINK-QA. The external inputs are indicated in dashed lines pointing towards the processes (rounded box) using them.

4.1 Components

Select This component is responsible for selecting the set of resources to be evaluated. This can be done through a variety of mechanisms including sampling the Web of Data, using a user specified set of resources, or looking at the set of resources to be linked by a link discovery algorithm. It is left to the user to decide whether the set of resources is a reasonable sample of the Data Network. **Construct** Once a set of resources is selected, the local network, as defined in Definition 2, is constructed for each resource. The local networks are created by querying the Web of Data. Practically, LINK-QA makes use of either SPARQL endpoints or data files to create the graph surrounding a resource. In particular, sampling is achieved by first sending a SPARQL query to a list of endpoints. If no data is found, LINK-QA falls back on de-referencing the resource. **Extend** The "Extend" component adds new edges that are provided as input to the framework. These input edges are added to each local network where they apply. Once these edges are added, we compute a set of new local networks around the original set of selected resources. The aim here is to measure the impact of these new edges on the overall Data Network. This impact assessment is done by the Compare component. **Analyse** Once the original local network and its extended local networks have been created, an analysis consisting of two parts is performed:

1. A set of metrics m is performed on each node v_i within each local network. This produces a set of metric results $m_i^{metric\ name}$ for each node v_i.
2. Metric results obtained per node are aggregated into a distribution. Note, that this distribution converges to the distribution of the overall Data Network as more resources are considered.

Compare The result coming from both analyses (before and after adding the new edges) are compared to ideal distributions for the different metrics. The comparison is provided to the user.

4.2 Implementation

The implementation is available as free software at http://bit.ly/Linked-QA, and takes as input a set of resources, information from the Web of Data (*i.e.* SPARQL endpoints and/or de-referencable resources) and a set of new triples to perform quality assessment on. The implementation is written in Java and uses Jena for interacting with RDF data. In particular, Jena TDB is used to cache resource descriptions. Any23 is used for dereferencing data in order to get a good coverage of possible publication formats.

The implementation generates HTML reports for the results of the quality assessment. These reports are divided in three sections:

1. An overview of the status of the different metrics based on the change of distance to the ideal distribution when the new links are added. The status is "green" if the distance to the ideal decreased and "red" otherwise. The relative change is also indicated. These statuses are derived from the change in $d^{metric\ name}$ observed when adding new links.
2. One graph per metric showing the distribution of the values for the different $m^{metric\ name}$ values obtained before and after adding the new set of links. The rendering of these graphs is done by the Google Graph API.
3. A table reporting for all of the metrics the resources for which the score $m_i^{metric\ name}$ has changed most after the introduction of the new links.

It is important to note that LINK-QA is aimed at analysing a set of links and providing insights to aid manual verification. There is no automated repair of the links nor an exact listing of the faulty links. Outliers - resources that rank farthest from the ideal distribution for a metric - are pointed out, but the final assessment is left to the user.

5 Metric Analysis

The framework is designed to analyse the potential impact of a set of link candidates prior to their publication on the Web of Data. To evaluate this, we test the links produced by a project using state of the art link generation tools: The European project LOD Around the Clock (LATC) aims to enable the use of the Linked Open Data cloud for research and business purposes. One goal of

the project is the publication of new high quality links. LATC created a set of linking specifications (link specs) for the Silk engine, whose output are link sets. In order to assess the correctness of link specs, samples taken from the generated links are manually checked. This results in two reference sets containing all the positive (correct, good) and negative (incorrect, bad) links of the sample. The link specs along with the link sets they produce, and the corresponding manually created reference sets are publicly available.[3] Based on these link sets we performed experiments to answer the following questions:

1. Do positive linksets decrease the distance to a metric's defined ideal, whereas negative ones increase it? If that is the case, it would allow us to distinguish between link sets having high and low ratios of bad links.
2. Is there a correlation between outliers and bad links? If so, resources that rank farthest from the ideal distribution of a metric would relate to incorrect links from/to them.

5.1 Impact of good and bad links

To try and answer the first question, we performed the following experiment: out of 160 link specifications created by LATC, we selected the 6 link sets (*i.e.* mappings) for which the manual verification of the links led to at least 50 correct and incorrect links. For each specification, we took separate random samples of 50 links from the postive and negative reference sets and ran LINK-QA. This was repeated ten times. Table 2 shows the aggregated results for each metric on positive and negative reference sets. The LATC link specification used to create the links are used as identifiers in the tables. The outcome of the ten runs is aggregated into three categories as follows: if no changes where detected in the results distributions, the category is "blank (B)"; a "C" is granted to link specification for which all the (in)correct were detected in at least half (5 in this case) of the runs. Least successful experiments are classified as "I".

A global success rate can be quickly drawn from Table 2 by considering the cumulative number of "C" and "I" to compute a recall score.

$$recall = \frac{\|I\| + \|C\|}{\|B\| + \|I\| + \|C\|} = \frac{21 + 20}{19 + 21 + 20} = 0.68$$

A precision index is given by the ratio between "C" and "I".

$$precision = \frac{\|C\|}{\|I\| + \|C\|} = \frac{20}{21 + 20} = 0.49$$

These two values indicate a mediocre success of our metrics on these data sets. From the table and these values, we conclude that common metrics such as centrality, clustering, and degree are insufficient for detecting quality. Additionally, while the Description Richness and Open SameAs Chain metrics look more

---

[3] https://github.com/LATC/24-7-platform/tree/master/link-specifications

Table 1. Detection result for each metric for both good and bad links. Blank - no detection, I - Incorrect detection, C - correct detection. (lgd = linkedgeodata).

	Centrality		Clustering		Degree		Description		SameAs	
	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad
linkedct-pubmed-disease			I	C	I	C	C	I	I	C
gho-linkedct-disease					C	I	C	I		
gho-linkedct-country			I	C	I	C				
geonames-lgd-island	C	C			I	I	C	C	C	C
gho-pubmed-country		I	I	I	I	I	C	I	I	C
geonames-lgd-mountain	C	I			I	C	C	I	I	C

Table 2. Detection result for each metric for both good and bad links. Blank - no detection. I - Incorrect detection C - correct detection.

promising, especially at detecting good and bad links, respectively, they report too many false positives for reference sets of the opposite polarity.

We now present a more in-depth analysis of the results found in the table focusing on the sensitivity of the metrics, their detection accuracy and their agreement.

Sensitivity of metrics The presence in Table 2 of blank fields indicates that the metric was not able to detect any change in the topology of the neighbourhood of resources, meaning that it fails at the first goal. We realise that the Degree metric is the only one to always detect changes. A behaviour that can be explained by the fact that adding a new link almost always yields new connections and thus alters the degree distribution.

The low performance of other metrics in detecting change can be explained by either a lack of information in the local neighbourhood or a stable change. The first may happen in the case of metrics such as the sameAs chains. If no sameAs relations are present in the local network of the two resources linked, there will be no chain modified and, thus, the metric will not detect any positive or negative effect for this new link. A stable change can happen if the link created does not impact the global distribution of the metric. The results found in Table 2 report changes in distributions with respect to the ideals defined, if the distribution does not change with the addition of the links the metrics is are ineffective.

Accuracy of detection With 21 "I" and 20 "C" in Table 2, we found as many incorrect results as correct ones. This result is unfortunately not good enough base decisions upon. There are a number of possible reasons for this low performance. It may be the case that network measures are not applicable at this level of network size. Indeed, a much larger network may be necessary for summarization effects to actually be applicable. Furthermore, the selected metrics may be inappropriate for Linked Data. Here, we enumerate 3 possible reasons.

1. **Definition of ideals**: the ideals are some target distribution we set as a universal goal Linked Data should aim for. It is however unclear whether

such a unique goal can be set for Linked Data. Our inspiration from social networks led us to aim at a small world topology, which does not correlate with the results found in our experiments; 2. **Coverage of sample**: the use of a sample of the studied network forces us to consider a proxy for the actual metrics we would have had computed on the actual network. Most noticeably, the centrality measure featured in our prototype is a rough approximation. For this metric in particular, a wider local neighbourhood around a resource would lead to better estimates. The same applies to the detection of sameAs chains which may span outside of the local neighbourhood we currently define; 3. **Validity of metrics**: the somewhat better performance of Linked Data specific network measures suggests that such tailored metrics may be a more effective than "class" metrics. The degree, clustering and centrality metrics look at the topology of the network without considering its semantics. However, as it is confirmed by our experiments, the creation of links is very much driven by these semantics and the eventual changes in topology do not provide us with enough insights alone. Our intuition, to be verified, is that effective metrics will leverage both the topological and semantic aspect of the network.

We believe a future path forward is to gain more empirical evidence for particular topologies and their connection to quality. The sampling of the Web of Data will also have to be reconsidered and may need to be defined with respect to a particular metric.

5.2 Detection of bad links

Our second research question is, whether LINK-QA can detect bad links in link sets with only a few bad links. Here, we are seeking a correlation between the ranking of outliers and the resources that are subjects of bad links. For this experiment, we took all LATC link specs with at least 50 positive and 10 negative reference links, and created samples of 45 positive and 5 negative links. LINK-QA was then run, and the process was repeated 5 times. Figure 3 shows the number of correct detections of outliers. With the exception of the cluster coefficient, the metrics show a bias for negative resources to be identified
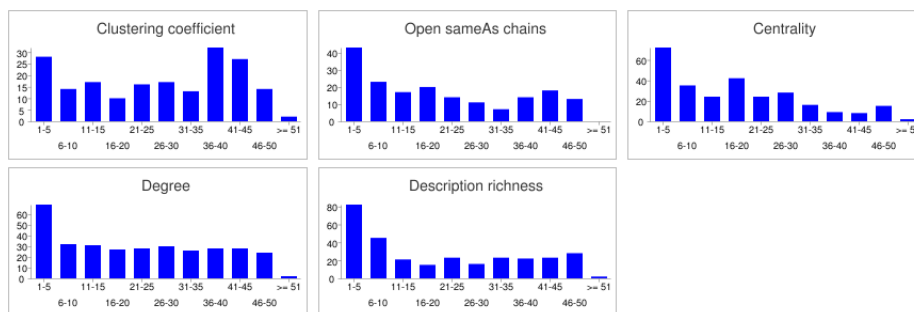

Fig. 3. Summary of outlier analysis. x-axis: rank of resources grouped in buckets of 5 (low values indicate outliers). y-axis: resource count.

as outliers. Although the remaining distributions do not seem directly suitable for detecting bad links, they show a trend in this direction, indicating that the predictions could be improved with the combination of multiple metrics. We exclude the cluster coefficient from the following analysis. Given a ranking of resources for each metric, we can assign each resource a sorted list of its ranks, e.g. $Fabulous\_Disaster \rightarrow (3, 3, 10, 17)$. A resource's n-th rank, considering n $= 1 \ldots 4$ metrics, is then determined by taking the $n - 1$-th element of this list. Ideally, we would like to see negative resources receiving smaller n-th ranks than positive ones. The distributions of the n-th ranks for all ns are shown in Figure 4. These charts indicate that a combination indeed improves the results: For example when combining 2 metrics, the probability of finding a negative resource on one of the first 5 ranks increases from about 20 to 30 percent, whereas an equi-distribution would only yield 10 percent (5 negative resources in 50 links). This effect increases, as can be observed in the right column: The positive-to-negative ratio is 0.6 for $n = 4$, which shows that a combination of metrics is effective in detecting incorrect links.
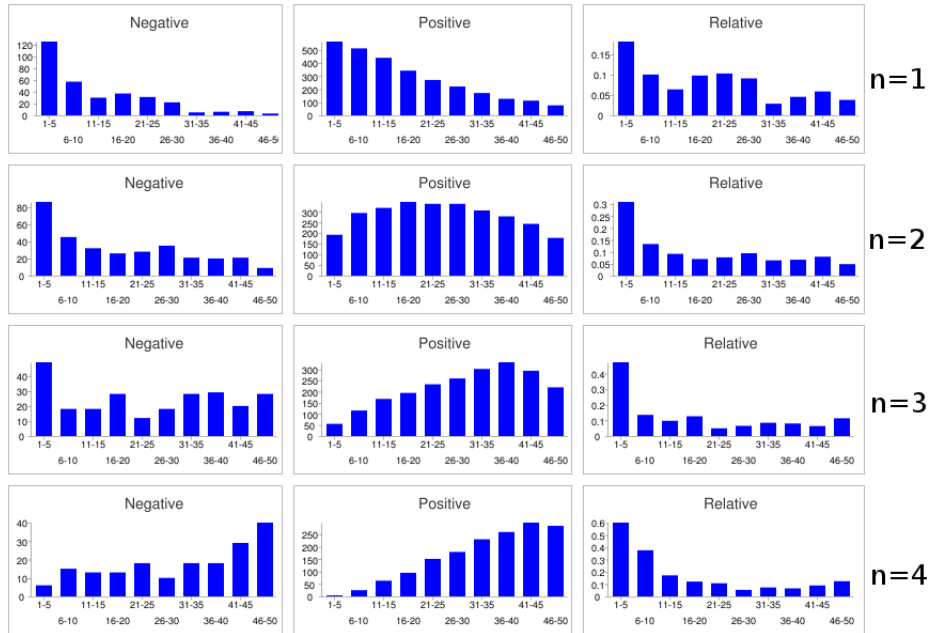

Fig. 4. Distribution of negative and positive resources by its n-th rank. For "negative" and "positive", the y-Axis shows the absolute number of resources detected for every bucket of five ranks. For "relative" it shows the ratio of negative to positive links.

6 Related work

In this section, we provide a review of related work touching on this paper. We particularly focus on quality with respect to the Semantic Web but also briefly touch on Network Analysis and the automated creation of links.

6.1 Quality

Improving data quality has become an increasingly pressing issue as the Web of Data grows. For example, the Pedantic Web group has encouraged data providers to follow best practices [15]. Much of the work related to quality has been on the application information quality assessment on the Semantic Web. In the WIQA framework [4], policies can be expressed to determine whether to trust a given information item based on both provenance and background information expressed as Named Graphs [5]. Hartig and Zhao follow a similar approach using annotated provenance graphs to perform quality assessment [19]. Harth *et al.*[13] introduce the notion of naming authority to rank data expressed in RDF based on network relationships and PageRank.

Trust is often thought as being synonymous with quality and has been widely studied including in artificial intelligence, the web and the Semantic Web. For a readable overview of trust research in artificial intelligence, we refer readers to Sabater and Sierra [20]. For a more specialized review of trust research as it pertains to the Web see [8]. Artz and Gil provide a review of trust tailored particularly to the Semantic Web [2]. Specific works include the IWTrust algorithm for question answering systems [24] and tSPARQL for querying trust values using SPARQL [14]. Our approach differs from these approaches in that it focuses on using network measures to determine quality.

Closer to our work, is the early work by Golbeck investigating trust networks in the Semantic Web [9]. This work introduced the notion of using network analysis type algorithms for determining trust or quality. However, this work focuses on trust from the point of view of social networks, not on networks in general. In some more recent work [11], network analysis has been used to study the robustness of the Web of Data. Our work differs in that it takes a wider view of quality beyond just robustness. The closest work is most likely the work by Bonatti *et al.*, which uses a variety of techniques for determining trust to perform robust reasoning [16]. In particular, they use a PageRank style algorithm to rank the quality of various sources while performing reasoning. Their work focuses on using these inputs for reasoning whereas LINK-QA specifically focuses on providing a quality analysis tool. Additionally, we provide for multiple measures for quality. Indeed, we see our work as complementary as it could provide input into the reasoning process.

6.2 Network analysis on the Web of Data

There are only a few studies so far about network analysis on the Web of Data, most of the significant existing work is focused on semantic schemas, paying a particular attention to either the schema relations [21] or the documents instantiating them [7]. Both studies show, on various datasets, that schemas tend to follow power-law distributions. Network analysis has also been used to rank results when searching for datasets on the Web of Data [22]. Our work applies these techniques to quality of the data published.

6.3 Automated creation of links

There is a large body of literature over the creation of links between data sets on the Web of Data. As a cornerstone of semantic interoperability, ontologies have attracted most of the attention over the last decade. Several ontologies mapping/integration/merging techniques, tools and platforms allows for the connection of different datasets on the schema level [6]. The Silk Link discovery framework [23] offers a more versatile approach allowing configurable decisions on semantic relationships between two entities. More recently, the LIMES [17] framework offers an efficient implementation of similar functionality. Driven by those approaches, there has been increasing interest in new ways to measure the quality of automated links. For example, Niu et al. propose confidence and stability as metrics for measuring link creation based on notions from the information retrieval literature [18].

Overall, our work sits at the convergence of the need for the quality assessment of the links automatically created and the use of network measures to perform that assessment.

7 Conclusion

In this paper, we described LINK-QA, an extensible framework for performing quality assessment on the Web of Data. We described five metrics that might be useful to determine quality of Linked Data. These metrics were analysed using a set of known good and bad quality links created using the mapping tool Silk. The metrics were shown to be partially effective at detecting such links. From these results, we conclude that more tailored network measures need to be developed or that such a network based approach may need a bigger sample than the one we introduced. We are currently looking at finding more semantics-based measures, such as the sameAs chains. We are also looking at the interplay of different measures and the combined interpretation of their results.

References

1. Adamic, L.A.: The small world web. In: 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL). pp. 443–452 (1999)
2. Artz, D., Gil, Y.: A survey of trust in computer science and the Semantic Web. J. Web Sem. 5(2), 58–71 (2007)
3. Barabási, A.L.: Linked. (Perseus, Cambridge, Massachusetts) (2002)
4. Bizer, C., Cyganiak, R.: Quality-driven information filtering using the WIQA policy framework. Journal of Web Semantics 7(1), 1–10 (Jan 2009)
5. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. International World Wide Web Conference (2005)

6. Choi, N., Song, I.Y., Han, H.: A survey on ontology mapping. ACM SIGMOD Record 35(3), 34–41 (2006)
7. Ding, L., Finin, T.: Characterizing the semantic web on the web. In: International Semantic Web Conference. pp. 242–257 (2006)
8. Golbeck, J.: Trust on the world wide web: a survey. Foundations and Trends in Web Science 1(2), 131–197 (2006)
9. Golbeck, J., Parsia, B., Hendler, J.A.: Trust networks on the semantic web. In: CIA. pp. 238–249 (2003)
10. Guéret, C., Groth, P., Harmelen, F.V., Schlobach, S.: Finding the Achilles Heel of the Web of Data : using network analysis for link-recommendation, vol. 6496, pp. 289–304. Springer (2010)
11. Guéret, C., Wang, S., Schlobach, S.: The web of data is a complex system - first insight into its multi-scale network properties. In: Proc. of the European Conference on Complex Systems. pp. 1–12 (2010)
12. Guéret, C., Wang, S., Groth, P., Scholbach, S.: Multi-scale analysis of the web of data: A challenge to the complex system's community. Advances in Complex Systems 14(04), 587 (2011)
13. Harth, A., Kinsella, S., Decker, S.: Using naming authority to rank data and ontologies for web search. The Semantic WebISWC 2009 2, 277–292 (2009)
14. Hartig, O.: Querying Trust in RDF Data with tSPARQL. In: Proc. of the 6th European Semantic Web Conference (ESWC). Heraklion, Greece
15. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the Pedantic Web. In: Linked Data on the Web Workshop (LDOW2010) at WWW'2010 (2010)
16. Hogan, A., Bonatti, P., Polleres, A., Sauro, L.: Robust and scalable linked data reasoning incorporating provenance and trust annotations. Journal of Web Semantics accepted (2011), to appear
17. Ngonga Ngomo, A.C., Auer, S.: Limes - a time-efficient approach for large-scale link discovery on the web of data. In: Proc. of IJCAI (2011)
18. Niu, X., Wang, H., Wu, G., Qi, G., Yu, Y.: Evaluating the stability and credibility of ontology matching methods. In: 8th Extended Semantic Web Conference (ESWC2011) (June 2011)
19. Olaf, H., Zhao, J.: Using Web Data Provenance for Quality Assessment. In: Proc. of the 1st Int. Workshop on the Role of Semantic Web in Provenance Management (SWPM) at ISWC. Washington, USA (2009)
20. Sabater, J., Sierra, C.: Review on Computational Trust and Reputation Models. Artificial Intelligence Review 24(1), 33 (2005)
21. Theoharis, Y., Tzitzikas, Y., Kotzinos, D., Christophides, V.: On graph features of semantic web schemas. IEEE Transactions on Knowledge and Data Engineering 20, 692–702 (2007)
22. Toupikov, N., Umbrich, J., Delbru, R., Hausenblas, M., Tummarello, G.: DING! Dataset Ranking using Formal Descriptions. In: WWW 2009 Workshop: Linked Data on the Web (LDOW2009). Madrid, Spain (2009)
23. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk: A link discovery framework for the web of data. In: 2nd Linked Data on the Web Workshop LDOW2009. pp. 1–6. CEUR-WS (2009)
24. Zaihrayeu, I., Pinheiro da Silva, P., McGuinness, D.L.: IWTrust: Improving User Trust in Answers from the Web. In: Proc. of 3rd International Conference on Trust Management (iTrust2005). pp. 384–392. Springer, Paris, France (2005)