How much navigable is the Web of Linked Data?*

Valeria Fionda[1], Enrico Malizia[2]

[1] Department of Mathematics, University of Calabria, Italy
[2] DIMES, University of Calabria, Italy

Abstract. Millions of RDF links connect data providers on the Web of Linked Data. Here, navigability is a key issue. This poster provides a preliminary quantitative analysis of this fundamental feature.

1 Motivation

Linked Data are published on the Web following the Linked Data principles [2]. One of them states that RDF links must be used to allow clients to navigate the Web of Linked Data from dataset to dataset. In particular, RDF links allow: *(i)* data publishers to connect the data they provide to data already on the Web; and *(ii)* clients to discover new knowledge by traversing links and retrieving data. Hence, navigability is a key feature. The Web of Linked Data is growing rapidly and both the set of data providers and the structure of RDF links continuously evolve. Is this growth taking place preserving the basic navigability principle?

In this poster we try to answer this question by analyzing the pay-level domain (PLD) networks extracted from the last three Billion Triple Challenge datasets. In addition, we also analyze the sameAs network obtained by considering only `owl:sameAs` links. Some recent works analyzed sameAs networks [1, 3] to provide some statistics on the deployment status and use of `owl:sameAs` links [3] and to evaluate their quality [1]. However, to the best of our knowledge, this is the first attempt to use the PLD and sameAs networks to perform a quantitative analysis of the navigability of the Web of Linked Data.

2 Methodology

Navigability indices. We model the Web of Linked Data as a directed graph $G = \langle V, E\rangle$ where $V = \{v_1, \ldots, v_n\}$ is the set of vertices and $E \subseteq V \times V$ is the set of edges. The vertices of G represent the pay-level domains that identify data publishers in the Web of Linked Data. The edges of G represent directed links between *different* pay-level domains (i.e., there are no loops) and are *ordered*

*V. Fionda's work was supported by the European Commission, the European Social Fund and the Calabria region. E. Malizia's work was supported by the ERC grant 246858 (DIADEM), while he was visiting the Department of Computer Science of the University of Oxford.

pairs of vertices (v_i, v_j), where v_i is the source vertex and v_j is the target one. Intuitively, an edge (v_i, v_j) models the fact that there is at least one URI having PLD v_i by dereferencing which an RDF link to a URI having PLD v_j is obtained.

We denote by $v_i \rightsquigarrow_G v_j$ the existence of a path from v_i to v_j in G (otherwise we write $v_i \not\rightsquigarrow_G v_j$). For a graph $G = \langle V, E\rangle$, $G^* = \langle V, E^*\rangle$ is the *closure* of G where $(v_i, v_j) \in E^*$ if and only if $v_i \neq v_j$ and $v_i \rightsquigarrow_G v_j$. We define the *reachability matrix* $R_G \in \{0,1\}^{n\times n}$ such that $R_G[i,j] = 1$ if and only if $(v_i, v_j) \in E^*$ (i.e., $v_i \rightsquigarrow_G v_j$). Moreover, we define the *distance matrix* $D_G \in \mathbb{N}^{n\times n}$ such that $D_G[i,j]$ is the length of the shortest path between v_i and v_j ($D_G[i,j] = \infty$ if $v_i \not\rightsquigarrow_G v_j$). When G is understood we simply write $v_i \rightsquigarrow v_j$, $R[i,j]$, and $D[i,j]$.

To evaluate the navigability of the Web of Linked Data we use two indices. The first one is the *reachability index* $\eta(G)$, corresponding to the *edge density* of G^*. The reachability index is the probability that between any two given vertices of G there exists a path. In particular, $\eta(G) = \frac{1}{n(n-1)} \sum_{v_i,v_j \in V, v_i \neq v_j} R[i,j] = \frac{|E^*|}{n(n-1)}$. This index takes into account only the reachability between vertices and implies that $\eta(G_1) = \eta(G_2)$ for any pair of graphs $G_1 = \langle V, E_1\rangle$ and $G_2 = \langle V, E_2\rangle$ such that $G_1^* = G_2^*$, even if $E_1 \subset E_2$ (or $E_2 \subset E_1$).

To take into account differences in graph topologies, we use the *efficiency index* $\widetilde{\eta}(G)$ [4]. This index exploits the distance matrix D to weight the reachability by the (inverse of the) length of the shortest path between vertices. Given a graph G, $\widetilde{\eta}(G) = \frac{1}{n(n-1)} \sum_{v_i,v_j \in V, v_i \neq v_j} \frac{R[i,j]}{D[i,j]}$, where $\frac{R[i,j]}{D[i,j]} = 0$ when $v_i \not\rightsquigarrow v_j$. It can be shown that for any graph G, $\widetilde{\eta}(G) \leq \eta(G)$, and given two graphs $G_1 = \langle V, E_1\rangle$ and $G_2 = \langle V, E_2\rangle$ such that $E_1 \subset E_2$ then $\eta(G_1) \leq \eta(G_2)$, while $\widetilde{\eta}(G_1) < \widetilde{\eta}(G_2)$. The index $\widetilde{\eta}(\cdot)$ has been used in literature to measure how efficiently small-world networks exchange information [4].

Intuitively, the closer $\eta(G)$ is to 1, the more G is similar to a strongly connected graph; on the other hand, the closer $\widetilde{\eta}(G)$ is to 1, the more G is similar to a complete graph. Note that, $\widetilde{\eta}$ combines information on reachability *and* information about the distances between pairs of vertices and it is not simply the arithmetic mean of the inverse of the shortest paths lengths.

Datasets. To perform our analysis we used the Billion Triple Challenge (BTC) datasets of 2010, 2011 and 2012[3]. Unfortunately, the BTC dataset for 2013 was not crawled and a dataset for 2014 was not available to the date of submission. We decided to use the pay-level domain (PLD) granularity to build our networks, where the PLD of a URI is a sub-domain (generally one level below) of a generic public top-level domain, for which users usually pay for. PLD domains in the Web of Linked Data are often in one-to-one correspondence with Linked Open Data datasets. We extracted the PLD network from each BTC dataset by considering each RDF quad and adding an edge between the PLD of the context URI and the PLD of the subject and object. In particular, we extracted two PLD networks: the first one (denoted by ALL) considers all types of links and the second one (denoted by SA) considers only `owl:sameAs` links.

[3] `http://km.aifb.kit.edu/projects/btc-X/`, `X={2010,2011,2012}`

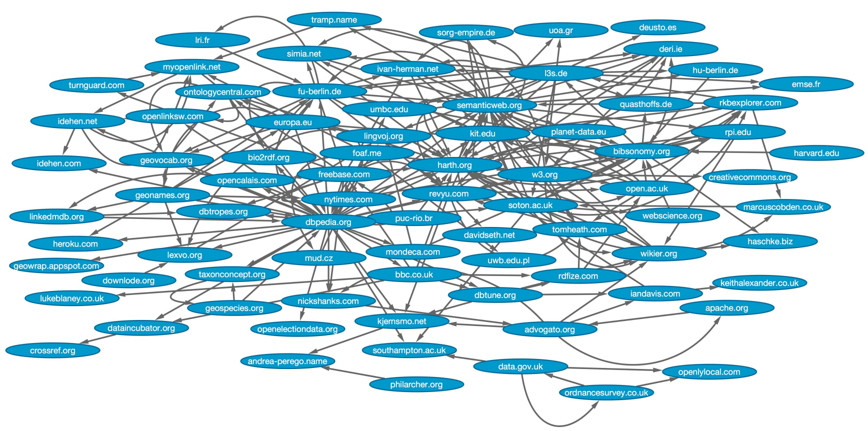

Fig. 1. The largest internal connected component of the PLD sameAs network extracted from the BTC2012 dataset.

Since BTC datasets are obtained by crawling the LOD cloud starting from a set of seed URIs, we also extracted from each network the largest internal connected component obtained by ignoring the PLDs "on the border" (i.e., those without any outgoing edge) that are probably those where the crawling stopped. We denote by ALL-I and SA-I are internal subnetworks extracted from ALL and SA, respectively. Fig. 1 shows the SA-I network of the BTC 2012 dataset.

3 Evaluation

Table 1 reports our results. The table shows that η and $\widetilde{\eta}$ on the complete network, for both ALL and SA, decrease in 2011 with respect to 2010 and are still decresing for the ALL network even in 2012. Moreover, the values obtained for both η and $\widetilde{\eta}$ are very small. For example, $\eta(\text{ALL}) = 4.899 \cdot 10^{-4}$ for the BTC 2012 dataset means that given a random pair of PLDs from the ALL network the probability that they are connected by a path is less than 0.5‰. Translated to the Web, this means that starting from a given a URI and following RDF links only a very small portion of the Web of Linked Data can be reached. However, an explanation for such a behavior could be that the BTC datasets are obtained by crawling the Web and it is reasonable to think that PLDs at the "border" of the network are those at which the crawling process stopped. If this is the case, some links can actually be missing and our indices can be biased. In general, a decrease over time of η and $\widetilde{\eta}$ on the full Web of Linked Data highlight a decrease in its navigability. Nevertheless, since in our case the BTC datasets are used as representative samples of the full Web of Linked Data, this decrease can be related to the fact that in 2012 and 2011 the crawler retrieved triples spanning more data providers than in 2010 and a large portion of them is on the

Network \ Index/Year	η			$\widetilde{\eta}$		
	2010	**2011**	**2012**	**2010**	**2011**	**2012**
All	0.034	0.002	$4.899 \cdot 10^{-4}$	0.009	$5.98 \cdot 10^{-4}$	$1.719 \cdot 10^{-4}$
SA	0.134	0.018	0.059	0.037	0.005	0.016
All-I	0.312	0.887	0.867	0.089	0.319	0.326
SA-I	0.400	0.658	0.497	0.131	0.223	0.187

Table 1. Summary of the analysis carried out.

border. Indeed, in general, both η and $\widetilde{\eta}$ decrease if the proportion of the nodes on the border increase with respect to the total size of the network.

For this reason we decided to analyze the internal largest connected components All-I and SA-I. As for All-I, on one hand it can be observed that the difference in the values of both indices between 2011 and 2012 is negligible. There is, on the other hand, a big increase in both indices for the 2011 network with respect to the 2010 one. It is evident, from these results, that the Web of Linked Data gained a lot in navigability from 2010 to 2011, according to the All-I sample, while the navigability remained almost unchanged in 2012. A similar trend can be identified also on the SA-I network, apart from a noticeable decrease in the navigability of the 2012 network compared to the 2011 one. Our results show that, for example, in 2012 given a random pair of PLDs from the All-I network the probability that they are connected by a path is greater than 86% with an efficiency of 0.326. It is worth to point out that, in a distributed environment as the Web, efficiency is a fundamental property that, besides reachability, is related to the number of dereferences that must be performed to move from a source PLD to a target one. Roughly speaking, lower values of efficiency for the same value of reachability translate in more traffic on the network.

4 Conclusions

Navigability is a key feature of the Web of Linked Data. We introduced some indices to quantitatively measure the connectivity of Linked Data providers and the efficiency of their connections. The results obtained show that, as hoped, the navigability of the Web of Linked Data is increasing with its growth. However, in order not to be biased toward a certain interpretation, it is important to stress that the results obtained could have been influenced by the crawling strategy used to build the BTC datasets used in our analysis. We plan to perform our analysis in the following years to monitor and hopefully confirm this trend.

References

1. G. Bartolomeo and S. Salsano. A spectrometry of linked data. In *LDOW*, 2012.
2. T. Berners-Lee. Linked data design issues, 2006.
3. L. Ding, J. Shinavier, Z. Shangguan, and D. McGuinness. SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl:sameAs in Linked Data. In *ISWC*, 2010.
4. V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87(19):198701, 2001.