Nogales Moyano, A., Sicilia, M.A. & García Barriocanal, E. 2022, "On the graph structure of the Web of Data", International Journal on Semantic Web & Information Systems, vol. 14, no. 2, pp 70-85.

Available at https://dx.doi.org/10.4018/IJSWIS.2018040104

(Article begins on next page)

1. Introduction

The "Linked Data approach" is the label for a set of techniques of publishing structured data on-line so that it can be interlinked and become more useful. Currently, Linked Data relies on standard Web technologies as the Hypertext Transfer Protocol[1] (HTTP), to retrieve data from the World Wide Web[2] (WWW) and the mechanisms of Uniform Resource Identifiers[3] (URI) for identification. It combines this with a rich representation of datasets based on Resource Description Framework[4] (RDF). The Web of Linked Data can thus be defined as a way to share open and structured data on the Web. Based on four principles exposed by Tim Berners-Lee in 2006[5], the objective is that of making the data more readable to computers by publishing it in RDF and using URIs to identify each resource. Finally data can be retrieved using the standardized language and protocol SPARQL[6] and RDF Query Language. These approaches can be framed in the field of the Semantic Web, Albert, Berners-Lee and Ficheti (1999) or Berners-Lee, Hendler and Lassila (1999). A general overview of what Linked Data is can be found in Bizer, Heath and Berners-Lee (2009).

The adoption of Linked Data best practices, described in Heath and Bizer (2011), has resulted in the publication of datasets by data providers in several domains. Datasets use RDF as language and expose triples to describe the information. Each triple has the structure of subject, predicate and object. Subjects are URIs representing resources. Objects could be URIs or particular values. Predicates are also represented by URIs and are the way to relate a subject with its object. The entities represented by the URIs can be looked up by using the HTTP protocol. In the case that the subject and the object of a triple belong to different datasets a link between them is created, it is what is often called a RDF link. If we take into account the different datasets and its RDF links, we can see the Web of Linked Data as a graph. Here the datasets will represent the nodes and the RDF links the edges between them. This structure of graph allows applications to navigate between them and discover new information. By adding new datasets, the Web of Linked Data (LOD) will evolve changing its structure. Since 2007 it has increased its size from a dozen of datasets to more than a thousand. A recent review can be found in Schmachtenberg, Bizer and Paulheim (2014).

The techniques used to analyze the Web as a graph can also be applied to the Web of Linked Data. If we study the link structure of LOD, different measures used in Network Analysis can be applied, Wasserman and Faust (1994): how to crawl information from it, importance of datasets by their sizes or understanding its behaviour looking at the evolution. This can help us answer questions as:

- What is the importance of different datasets in the network structure?
- Are there interesting connected component?
- Can the structure of the Web of Linked Data help us?

This paper is reporting on the main empirical findings on a comprehensive analysis of the graph structure of the Web of Linked Data. Following approaches used successfully in the past to analyse that structure for the Web of documents. This led to important insights helping to innovation in search engines or an increased understanding of the social structure of Web communities. First of all we are obtaining some metrics about the datasets in LOD and its usage. Then we will apply Social Network Analysis (SNA) techniques giving us a general picture of the Web of Linked Data.

Results show that the structure of the Web is very compact which a low distance between nodes. The nodes also have a reasonable number of edges and most of them are close to the rest of the graph. Finally we can conclude that the structure of the Web fits with the theory of the bow-tie.

The rest of this paper is structured as follows. Section 2 provides a brief overview of existing studies on the graph structure of the Web and its potential usefulness to be applied to the Web of Linked Data. Section 3 outlines the main objectives of the study presented here. Then, materials and methods are described in Section 4 and results are discussed in Section 5. Finally, conclusions and outlook are presented in Section 6.

2. Background

SNA has been successfully performed in order to know the structure and also to measure the classical Web of Documents, also the Web of Data. The following papers have as aim give measures or the structure of a set of any of these networks.

If we talk about the Web in general, the first studies about its structure can be found in Barabási and Albert (1999) working with 325 thousand webpages from nd.edu. Later, Kumar et al (1999) made an experiment obtaining communities over a crawl of more than 200 million of pages. During this year, also Albert, Jeong and Barabási (1999), reported a study about the diameter of the WWW. In Serrano et al (2000), four different Web graphs were obtained using four different crawlers. The analysis brings several differences both in the quantitative and the qualitative aspects. The evolution of the Web is studied in Hall and Tiropanis (2012), in the paper we found the definition of Web Science. An important study about the structure of the Web was reported by Broder et al (2000). They analysed 200 million of pages and 1.5 billion links. It concluded that the World Wide Web can be grouped in four big sets with the shape of a bow-tie. Contrasting the structure described in this paper, we have Donato et al (2008), proposing that the structure is similar to a daisy. The structure of a teapot is found in Zhu et al (2008), where the Chinese Web is analysed. We have also to review, Meusel et al (2014), confirming that the structure has not changed a lot and it maintains the shape of a bow-tie. In this paper the crawl, provided by the Common Crawl[7] project, has 3.83 billion of documents. The two papers, Broder et al (2000) and Meusel et al (2014), will be used as the main point of departure for ours, provided that we will work with the Web of Linked Data. In our paper we are applying some of the techniques exposed in these papers but the structure we are analysing is the Web of Linked Data.

There are some works analysing the structure of little sets in the Web. For example, Boldi et al (2002) makes an analysis of the structure of the African Web. In Baeza-Yates and Poblete (2003) the evolution of the Chilean Web between 2000 and 2002 is presented. This paper is contrasted with Baeza-Yates and Poblete (2006) where a similar study is made, concluding that there are many stable Websites but also chaotic changes. In Somboonwi, Suzuki and Kitsuregawa (2008) the evolution of the structure of the Thai web has been analysed from 2004 to 2007. Comparing with the previous paper, the object of study in our paper will be the whole Web of Linked Data.

There are several papers in which some metrics applying SNA are studied in the field of Linked Data. The following papers analyse only a few properties and work with little sets of LOD. A study of the connectivity of a crawl can be found in Ge et al (2010), providing metrics like diameter and density. Another property like semantic distance is measured in Passant (2010) using DBPedia[8] as use case. The usage of SNA applied to two ontologies is shown in Hoser et al (2006), measuring their betweenness and eigenvector centrality. A big analysis of the Semantic Web is done in Ding and Finin (2006) giving some global properties like size. Also it is demonstrated that the distribution or complexity follows a power law distribution. The study of network characteristics like degree, connectivity and reachability of a big data set, containing about 9 millions of RDF document, is reported by Cheng and Qu (2008). These papers give some NSA metrics of little sets in the Web of Linked Data; in our paper we are reporting not only a few characteristics and we are studying the whole Web of Linked Data.

Some deeper analysis, trying to give metrics or a general structure of the Web of Linked Data, have been reported in recent years. The first measure of LOD was made by Hausenblas et al (2012), in which the size of the Linking Open Data Project[9] is measured. Then, in Rodriguez (2009) a graph analysis of the Linked Data is done giving some structural properties like diameter or average path length. It also gives a structural analysis and concludes that the graph obtained is not strongly connected. In Rietveld and Hoekstra (2013) the usage of datasets of the Linked Data Cloud are studied using YASGUI[10], a web-based query editor. One of the most famous papers giving metrics of LOD is Demter et al (2012). Here the project LODStats[11] is presented in which we can find the usage of datasets or vocabularies in the Web of Linked Data. Dividino et al (2014) analysed the evolution of datasets used in LOD. None of these papers covers the amount of metrics presented in ours. Also most of the analysis has been made in previous years so the size of the Web of Linked Data has increased.

Finally, in Schmachtenber, Bizer and Paulheim (2014), we find an overview of the linkage relationships between datasets in the form of an updated LOD cloud diagram. The crawl provided by this paper, which is the most complete we could find, will be used in this paper. In this paper two NSA metrics are given the degree distribution and the connected components. In our case the analysis will be more complete as it has new metrics like diameter, centrality or the analysis of structure as a bow tie. In the case of the metrics that were calculated in the previous paper like degree or the connected components. We have added histograms and calculated power laws for the degree distribution and the size of the different components in the case of the connectivity.

Taking into account the previous works, our contribution will consist of making a more general analysis of the Web of Linked Data. First of all, the most recent paper reporting analysis was done in 2009, from that time till now the use of Linked Data has significantly grown. We are also including most of the graph patterns like diameter or centrality. Finally we are applying some principles of Social Network Analysis obtaining characteristics like degree distribution or connectivity. Taking into account all that information a general structure of the graph will be obtained.

3. Objectives

We will work will work with the directed graph formed by triples. The nodes of the graph correspond to the datasets and the edges represent the RDF links. First of all we will obtain general statistics as size of the datasets or metrics like diameter. Then we will study properties like degree distribution or connected components. There are several reasons for developing an understanding of the graph structure of the Web of Linked Data, including the following:

- What is the importance of different datasets in the network structure?
- Are there interesting connected component?
- Can the structure of the Web of Linked Data help us?

The study of the points presented above has had a big impact in several papers as the following. Isele et al (2010) presents LDSpider[12], an open-source web crawling framework allowing to move through the Web of Linked Data to download it. In Hogan, Harth and Decker (2007) the famous web algorithms PageRank[13] and HITS[14] are adapted to work over RDF Graphs. A linked data aggregation framework that solves individual conflicts and measures the quality of the data aggregated, is described in Knap and Michelfeit (2012). IRLBot[15] is presented in Lee et al (2008), this crawler has a set of web-crawler algorithms solving problems that arisen when other algorithms tried to work with big amounts of URLs. Finally Alsarem (2013) presents a technique that transforms RDF graphs into bipartite graphs representations applying ranking algorithms to produce enhanced snippets.

4. Objectives

Having a snapshot of the Web of Linked Data is difficult nowadays as the number of datasets available grows diary, and traditional crawling mechanisms are not guaranteed to produce valid samples. The production of a dataset is not similar to the creation of pages on the Web, as it is tied today to particular institutions and projects that are capable of deploying the technology and that have an interest in enabling innovations. For that reason, in this paper the departure data has been the subset of linked datasets that are supposed to belong to the LOD Cloud initiative[16].

4.1. Data gathering

For the purpose of the present study, we have used the dump file provided by the University of Mannhein[17] also used in Schmachtenberg, Bizer and Paulheim (2014). The dump has a size of 42.68 Gigabytes uncompressed. It contains 188,440,372 n-quads distributed in 1,014 datasets. This dumps includes datasets from lod-cloud group in the datahub.io dataset catalog[18], plus the Billion Triple Challenge 2012[19] and datasets advertised on the public-lod@w3.org mailing list[20] since 2011.

4.2. Dump, normalization and cleaning

The crawl is formed by n-quads with the format: subject, predicate, object and dataset. Having this structure, we are interested in working with it as a csv file. For that purpose we need the data to be normalized and cleaned. First of all we have to set the structure of the file so we can have only four columns separated by blanks. In order to do that we have to change the blanks found in some URIs for underscores so they could not be confused with the separators of the columns. The final file will only have the three blanks that separate the four columns of each n-quad. Then we have realized that some URIs are ill-formed, using Hexadecimal notation for special characters like _:httpx3Ax2Fx2Fwwwx2Efabianabel, we are skipping this instances which are the 7.32% of the dump. We have also found instances where the subject or the object belongs to a regular Website instead of a Linked Data dataset, this instances has also not been taking into account. Finally, 166 of the datasets could not be crawled, so they will not be part of the final file.

4.3. Network analysis

The analysis of the file has been developed using IPython notebooks. First of all we have to create our graph. The nodes have been added using a file that contains the name of all datasets. The information to create the edges has been obtained from the file that we have normalized and cleaned before. Then we have used the typical iPython libraries like numpy[21] and pandas[22] to manage the data and matplotliB[23] to present the information with graphics. We also have used

NetworkX[24] to create the graph and get SNA measures. We also have obtained some power laws using the poweRlaw[25] library, presented in Ludbrook et al (2014).

5. Results and discussion

As we said before we are studying the LOD as a graph. A formal definition is given bellow, Passant (2010).

Definition 1. A dataset following the Linked Data principles is a graph G such as $G = (R, L, I)$ in which $R = \{r1 , r2 , \ldots, rn \}$ is a set of resources — identified by their URI —, $L = \{l1, l2, \ldots, ln\}$ is a set of typed links — identified by their URI — and $I = \{i1, i2, \ldots, in\}$ is a set of instances of these links between resources, such as $ii = \langle lj, ra, rb\rangle$

Scaling to the Web, the Linking Open Data cloud is then defined as the union of all the graphs Gi that are published (and interlinked) on the Web, i.e. $LOD = \cup i\ Gi$.

In other words we are studying a graph whose nodes will be datasets and whose edges will be the URIs interlinking each dataset. A dataset will only correspond to one dataset but two datasets can be connected with more than one edge.

Taking into account this definition applied to our dump. We will only take into account the instances that connect two datasets. Using the directed graph obtained, we are studying some characteristics to discover some general conclusions about the structure of the graph. We have chosen some main characteristics, which are enumerated and defined below. The first subsection will give us some general metrics of the whole graph and the size of the datasets. Subsection two will show the classification of the datasets according to its degree. The following subsection talks about the connectivity of the different datasets. Finally we have a subsection trying to know if LOD follows a particular structure like the bow-tie.

5.1. Overall structure

Using the graph given before it is possible to set some general measures. First of all we have to take into account that LOD is a disconnected graph, so some measures like average path length cannot be computed. Table 1 shows a summary of the graph. The definitions of the metrics presented in the table are shown below.

Table 1. General statistics of LOD.

Statistic	Value
Number of vertices	1,014
Number of edges	4,692
Strongly connected	False
Weakly connected	False
Diameter	9
Degree centrality	0.0019
Closeness centrality	0.12

A directed graph is strongly connected if there is a path between all the pairs of nodes. If we have a maximal strongly connected subgraph, we can consider it a Strongly Connected Component (SCC) of a graph. NetworkX uses Tarjan's algorithm, Tarjan (1972), with Nuutila's modifications, Nuutila and Soisalon-Soininen (1999), to find SCCs in a graph. In our case, LOD is not strongly connected.

A weakly connected graph is when we avoid the directions of the edge and we obtain a strongly connected graph. In this case as some nodes are disconnected, the graph could not bet weakly connected. Although our graph is not strongly connected, neither weakly connected, studying the existence of connected components is relevant. A connected component is a subgraph in which any two nodes are connected. A more exhaustive analysis of this will be made later.

For diameter, there are a few definitions to consider what is the diameter of a graph. We are using the one proposed in Tauro et al (2001). Effective diameter or eccentricity is the minimum number of hops in which some fraction (say, 90%) of all connected pairs of nodes can reach each other. As the diameter here is very low regarding the total amount of nodes, it indicates that they are in proximity and the graph is compact.

There is also an importance in measuring centrality. In this case we have obtained the degree and closeness centrality. Using the definitions provided by Opsahl, Agneessens and Skvoretz (2010). Degree centrality, taking into account that the degree of a node is its number of connections, was computed as the number of ties or neighbours of a node. Closeness was the inverse of the sum of all shortest paths to others or the smallest number of ties to go through to reach all others individually. The low degree centrality lets us know that datasets are only linked to a few. In the case of the closeness centrality indicates that most of the nodes are close to the rest of the nodes of the graph.

Another interesting measure of LOD consists of studying the distribution of the information. As this is formed by datasets we are interested in knowing which have the biggest size. Taking into account the number of n-quads instances of each dataset. The following table shows us the top 5 datasets ordered by size. For each dataset name, number of occurrences and its percentage regarding the total size of the Web of Linked Data are indicated. The quartiles of this distribution have the following values: Q_1 is 44, Q_2 is 259 and Q_3 is 3569.75.

Table 2. Use of LOV vocabularies in datasets.

Dataset	Number of occurrences	Percentage
opendata.euskadi.net/	81,162,382	43.07 %
fr.dbpedia.org	13,767,913	7.3 %
dbpedia.org	8,130,084	4.31 %
dbtropes.org	6,930,857	3.67 %
estatwrap.ontologycentral.com	5,665,528	3.006 %

Here we find datasets like Open Data Euskadi[26], a catalogue of public data from the government of Euskadi. In the second position we find the dataset of the Dbpedia in French[27]. The following dataset in the table is for Dbpedia[28], a project about extracting structured data from the Wikipedia, Lehman (2012). DBTropes[29] is a linked data wrapper for the TVTropes.org community. Finally Linked Eurostat (Ontology Central)[30] is a mediator that translates original Eurostat files to RDF.

5.2. Degree distribution

Another interesting measure is that related with the degree of nodes. Considering the whole graph, the probability distribution of its degrees, is what we call degree distribution. One of the reasons why we are interested in this characteristic is to know if the Web of Linked Data is distributed according to some known distribution functions.

One of the experiments consists of obtaining how the in-degree and out-degree are distributed according to power laws as many other phenomena in the Web, Faloutsos, Faloutsos and Faloutsos (1999). The in-degree of a node is the summary of the edges reaching the node and the out-degree is the summary of the edges leaving the node. We define a power law as a mathematical relation between two quantities where one of the quantities decreases while the other increases. In our case we are measuring the number of datasets against its degree. To calculate that we are using the mathematical definition of a power-law distribution is as it follows.

Definition 2:

$$p(x) = Cx^{-\alpha}\ for\ x \geq x_{min} \tag{2}$$

In this definition, x corresponds to a range of values, and C and α are constants. In fact, C is derived as $C = (\alpha - 1)x_{min}^{\alpha}$. We have to take into account that α> 1 is a requirement for a power-law form to normalize. A histogram showing the distribution of the in degree is shown in figure 1.

Figure 1. **In degree histogram.**

A graphic proof that the in degree distribution fits with a power law can be seen in figure 2.

Figure 2. **Layout and margin requirements for tables and figures.**

Table 3. Top in-degree datasets.

In-degree	Dataset
331	w3.org
226	dbpedia.org
132	reference.data.gov.uk
131	geonames.org
89	semanlink.net

In this table we have new datasets like WordNet 2.0 (W3C)[31], which presents a standard conversion of Princeton WordNet to RDF. The third position is for linked UK government data. Semanlink is a dataset that can be used to manage files, bookmarks and short text notes, Servant (2006). The last dataset contains Social web content (Personal profiles/posts) from http://ldodds.com. Datasets with high in degree tend to be more important as they are supposed to be more popular

In the case of the out-degree, x_{min}=27 and α=2.53, so it follows a power law. With this information we have built table 5 and 6. The histogram of this distribution is show in figure 3.

Figure 3. **Out degree histogram.**

Also a figure showing how the out degree distribution fits with a power law can be seen in figure 4. If we compare it with figure 2 it shows that the in degree distribution fits better than the out degree.

Figure 4. **Out degree histogram**

Table 4. Top out-degree dataset.

Out-degree	Datasets
177	dbpedia.org
112	semanticweb.org
96	data.semanticweb.org
92	bibsonomy.org
90	fr.dbpedia.org

New datasets not reviewed before can be found in the table below. For example in the second position we have the Semantic Wiki about the Semantic Web community[32]. The following dataset is called Semantic Web Dog Food[33] and here you can find information about the main conferences and workshops in the area of Semantic Web research. Finally there is a dataset called BibSonomy[34], which pertains to a system for sharing bookmarks and lists of literature. Datasets with high out-degree can disperse information faster so they are considered as influential.

5.3. Connectivity

Connectivity, in graph theory it is defined as the minimum number of elements (nodes or edges) which need to be removed to disconnect the remaining nodes from each other. The connectivity allows us to define the structure of the graph; in which sets are grouped the nodes and how they are connected between them. A deeper explanation of the structure can be found in section 5.4.

There are two main measures related with connectivity: strongly connected components and weakly connected components that have been defined before. In both cases we are interested in finding the sizes of the components. In the case of the strongly connected components. We have one of 511 nodes, another of 3, one of 8, three formed by a pair of nodes and the remaining 486 have only one. If we talk about the weakly connected components. There is one of 904 nodes and 110 formed by one node.

5.4. Structure of the Web of Linked Data

Considering the results obtained in the experiments related to the connectivity of the graph, we tried to know if it complies the bow-tie theory exposed in Broder et al (2000). According to this paper the structure has six components:

- SCC, the strongly connected component that is the core of the structure.
- IN, is made by datasets that can reach the SCC component but cannot be reached.
- OUT, similar to the IN component but formed by datasets that are reached from the SCC component.
- TUBES, has nodes that are not in the SCC component, are reachable from IN and can reach OUT.
- TENDRILS, are datasets that can not reach and are not reachable from SCC, but belong to IN or OUT components.
- DISCONNECTED, datasets that has no connections. It cannot be consider as a real component of the structure.

To compute if the graph follows his structure, we are using Pajek. Pajek is a program for analysing and visualizing large networks, Batagelj and Mrvar (1999). Table 7 shows how many nodes have each component after analysing the graph with Pajek.

Table 5. Bow-tie components.

Components	Nodes
SCC	511
IN	283
OUT	101
TUBES	0
TENDRILS	9
DISCONNECTED	110

A graph generated by Pajek is shown in figure 5. Here we can see the different components grouped with its nodes drawn with different colours. Yellow is for SCC, green for IN, red for OUT, purple for TENDRILS and blue for DISCONNECTED. Being part of the core, which is the SCC component, we can highlights datasets like WordNet 2.0, Dbpedia or Semantic Wiki. These datasets will be taken into account when designing crawl strategies as they will be in the center of the structure.

Figure 5. **Bow-tie structure.**

5.5. Connectivity

One fundamental limitation of the conclusions provided here is that they are based on the dump provided by Mannheim University. We have to take into account that maybe some datasets could not been crawled. From the 1014 datasets that were supposed to be crawled, 166 gave us no data. We have also discovered that sometimes n-quads are bad formatted making difficult to know which datasets do they belong.

5.6. Limitations

One fundamental limitation of the conclusions provided here is that they are based on the dump provided by Mannheim University. We have to take into account that maybe some datasets could not been crawled. From the 1014 datasets that were supposed to be crawled, 166 gave us no data. We have also discovered that sometimes n-quads are bad formatted making difficult to know which datasets do they belong.

6. Conclusions and outlook

The main target of this paper was to give some characteristics and a global view of the structure of the Web of Linked Data. First of all measures like diameter and closeness and degree centrality tell us that LOD is a compact structure in which the distance between nodes is low. We have also demonstrated the both the in and out-degree follow a power low. That means that most of the nodes have a reasonable amount of edges leaving or reaching the node. Regarding the datasets we can highlight that the Open Data Euskadi is the biggest one. In the case of the datasets with more connections to other datasets, WordNet 2.0 and Dbpedia will be on the top. Finally we have checked that the structure of the Web of Linked Data accomplished the theory of the bow-tie, so we can organized it into five different components with different characteristics.

All the characteristics found could be applied in the design of crawl strategies. Another discovery, which can contribute the objective defining in section 3, is the information of the datasets that can be used in rankings. The other two objectives understanding and predicting the evolution need studies in different moments so we could know how it has changed.

The work reported in this paper can be used in the future to study different tasks. We could be interested in studying in a deeper way the different components of the structure. We could make another analyse in a few years to conclude how the Wed of Linked Data evolves. It can also be applied to develop tools aimed to work with the data stored in the different datasets.

Notes

1.http://www.w3.org/Protocols/
2. http://www.w3.org/
3. http://www.w3.org/Addressing/URL/uri-spec.html
4. http://www.w3.org/RDF/
5. http://www. w3.org/DesignIssues/LinkedData.html
6. http://www.w3.org/TR/rdf-sparql-query/
7. http://commoncrawl.org/
8. http://www.dbpedia.org
9. http://www.linkeddata.org
10. http://www.yasgui.laurensrietveld.nl/
11. http://stats.lod2.eu/
12. https://code.google.com/p/ldspider/
13. http://en.wikipedia.org/wiki/PageRank
14. http://en.wikipedia.org/wiki/HITS_algorithm
15. http://irl.cs.tamu.edu/crawler/
16. http://lod-cloud.net/
17. http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/
18. http://datahub.io/group/lodcloud
19. http://km.aifb.kit.edu/projects/btc-2012/
20. http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/
21. http://www.numpy.org/
22. http://pandas.pydata.org/
23. http://matplotlib.org/
24. http://networkx.github.io/
25.https://pypi.python.org/pypi/powerlaw
26. http://opendata.euskadi.net/que-es-opendata/
27. http://fr.dbpedia.org/
28. http://dbpedia.org
29. http://datahub.io/es/dataset/dbtropes
30. http://estatwrap.ontologycentral.com/
31. http://w3.org
32. http://semanticweb.org
33. http://data.semanticweb.org
34. http://bibsonomy.org

Funding

The work leading to these results has received funding from the European Union Seventh Frame-work Programme, in the context of the SemaGrow (ICT-318497), http://www.semagrow.eu/

References

1 Albert R, Jeong H and Barabási A. Internet: Diameter of the world-wide web. *Nature* 1999; 401: 130-131.

2 Alsarem M. A Generic Approach Based on Linked Data to Enhance Web Information Retrieval and Increase User Satisfaction. In Proceedings Dans *la COnférence en Recherche d'Information et Applications*, Nauchatel, 2013, pp. 299-304.

3 Baeza-Yates R and Poblete B. Evolution of the Chilean Web structure composition. In *Proceedings of Latin American Web Conference*, Santiago, 2003, pp. 11-13.

4 Baeza-Yates R and Poblete B. In *Proceedings of the 3rd International Workshop on Web Dynamics*, New York, 2006, pp. 1464-1473.

5 Barabási AL and Albert R. Emergence of Scaling in Random Networks. *Science* 1999, 286: 509-512.

6 Batagelj V and Mrvar A. Pajek - Analysis and Visualization of Large Networks. Berlin: Springer, 1999.

7 Berners-Lee T and Fischetti M. Weaving the web: The original design and ultimate destiny of the World Wide Web by its inventor. New York: Harper Collins Publishers, 1999.

8 Berners-Lee T, Hendler J and Lassila O. The Semantic Web. *Scientific American* 2001; 284: 34-43.

9 Bizer C, Heath T and Berners-Lee T. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems* 2009; 5: 1-22.

10 Boldi P, Codenotti B, Santini M and Vigna S. Structural properties of the African Web. In *Proceedings of the eleventh international conference on World Wide Web*, Honolulu, 2002.

11 Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A and Wiener J. Graph structure in the Web. *Computer Networks* 2000; 33: 309-320.

12 Cheng G and Qu Y. Term Dependence on the Semantic Web. *In Proceedings of the 7th International Conference on The Semantic Web*, 2008, Karlsruhe, pp. 665-680.

13 Demter J, Auer S, Martin M and Lehmann J. LODStats---An Extensible Framework for High-performance Dataset Analytics. In *Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management,* 2012, Galway, pp. 353-362.

14 Ding L and Finin T. Characterizing the Semantic Web on the web. In *Proceedings of the 5th International Semantic Web Conference*, 2006, Athens, pp. 242-257.

15 Dividino R, Gottron T, Scherp A and Gröner G. From Changes to Dynamics: Dynamics Analysis of Linked Open Data Sources. In *Proceedings of the Workshop on Dataset ProfiIling and Federated Search for Linked Data,* 2014, Crete.

16 Donato, D., Leonardi, S., Millozzi, S. and Tsaparas, P. Mining the inner structure of the web graph. *Journal of Physics A: Mathematical and Theoretical* 2008; 41, 224017.

17 Faloutsos M, Faloutsos P and Faloutsos C. On power-law relationships of the Internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, Cambridge, 1999, pp. 251-262.

18 Ge W, Chen J, Hu W and Qu Y. In *Proceedings of the 7th international conference on The Semantic Web: research and Application,* 2010, Heraklion, pp. 257-271.

19 Hall W and Tiropanis T. Web evolution and Web Science. *Computer Networks* 2012; 56: 3859-3865.

20 Hausenblas M, Halb W, Raimond Y and Heath T. What is the Size of the Semantic Web? In *Proceedings of the International Conference on Semantic Systems,* 2008, Graz.

21 Heath T and Bizer C. (2011). Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers, 2011.

22 Hogan A, Harth A and Decker S. Performing Object Consolidation on the Semantic Web Data Graph, In *Proceedings of I3: Identity, Identifiers, Identification*, 2007, Banff.

23 Hoser B, Hotho A, Jäschke R, Schmitz C and Stumme G. Semantic Network Analysis of Ontologies. In *Proceedings of the 3rd European Semantic Web Conference,* 2006, Budva, pp. 514-529.

24 Isele R, Umbrich J, Bizer C and Harth A. LDspider: An Open-source Crawling Framework for the Web of Linked Data.. In *Proceedings of 9th International Semantic Web Conference*, 2010, Shanghái.

25 Knap T and Michelfeit J. Linked Data Aggregation Algorithm: Increasing Completeness and Consistency of Data. In *Proceedings IEEE 36th Annual Computer Software and Applications Conference Workshops*, 2012, Izmir.

26 Kumar R, Raghavan P, Rajagopalan S and Tomkins A. Trawling the Web for emerging cyber-communities. In *Proceedings of the eighth international conference on World Wide Web,* New York, 1999, pp. 1481-1493.

27 Lee HT, Leonard D, Wang X and Loguinov D. IRLbot: scaling to 6 billion pages and beyond. In *Proceeding of the 17th international conference on World Wide Web,* New York, 2008, pp. 427-436.

28 Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, van Kleef P, Auer S and Bizer C. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal 2012; 6: 167-195.

29 Ludbrook J, Alstott J, Bullmore E and Plenz D. powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1): e85777. 2014.

30 Meusel R, Vigna S, Lehmberg O and Bizer C. Graph structure in the web --- revisited: a trick of the heavy tail. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion,* Geneve, 2014, pp. 427-432.

31 Nuutila E and Soisalon-Soininen E. On Finding the Strongly Connected Components in a Directed Graph. Information Processing Letters 1999; 49: 9-14.

32 Opsahl T, Agneessens F and Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 2010; 32: 245-251.

33 Passant A. Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, 2010.

34 Rietveld L. and Hoekstra R. YASGUI: Not Just Another SPARQL Client. In *Lecture Notes in Computer Science*, 2013, Montpellier, pp. 78-86.

35 Rodriguez MA. A Graph Analysis of the Linked Data Cloud. CoRR, 2009.

36 Schmachtenberg M, Bizer C and Paulheim H. Adoption of the Linked Data Best Practices in Different Topical Domains. In *Lecture Notes in Computer Science The Semantic Web, ISWC 2014*, Riva del Garda, 2014, pp. 245-260.

37 Serrano MÁ, Maguitman AG, Boguñá M, Fortunato S and Vespignani A. Decoding the structure of the WWW: A comparative analysis of Web crawls. *Computer Networks: The International Journal of Computer and Telecommunications Networking* 2000; 33: 309-320.

38 Servant FP. Semanlink. Jena User Conference (JUC), 2006.

39 Somboonviwat K, Suzuki S and Kitsuregawa M. Connectivity of the Thai Web Graph. In *Proceedings of the 10th Asia-Pacific web conference on Progress in WWW research and development* ., Shenyangm 2008, pp. 613-624.

40 Tarjan RE. Depth-First Search and Linear Graph Algorithms. SIAM Journal on Computing 1972; 146-160.

41 Tauro L, Palmer C, Siganos G and Faloutsos M. A simple conceptual model for the Internet topology In *Proceedings of Global Internet*, San Antonio, 2001, pp. 1667–1671.

42 Wasserman S and Faust K. Social network analysis: methods and applications. Cambridge; New York: Cambridge University Press, 1994.

43 Zhu JJH, Meng T, Xie Z, Li G and Li X. A teapot graph and its hierarchical structure of the Chinese web. In *Proceedings of the 17th international conference on World Wide Web,* New York, 2008, pp. 1133-1134.