Topological properties of the Linked Open Data Webgraph

Tope Omitola[1], Ian Millard[1], Hugh Glaser[1], Nicholas Gibbins[1], and Nigel Shadbolt[1]

Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
{ tobo, icm, hg, nmg, nrs }@ecs.soton.ac.uk

Abstract. In this paper, we present an experimental study of the connectivity properties of the Linked Open Web of Data. Using these connectivity properties, we attempt to discern structure out of the basic elements, i.e. the datasets, and the linkages between them. We report our experimental findings of its topological properties, such as its degree centrality, geodesics, prestige, and its density. Understanding the Open Data Webgraph structure is important for a number of applications such as effective linked data consumption, federated querying, querying optimisation, and reachability analysis.

1 Introduction

Space is very important in our lives. We live and interact in space, and our lives are rooted and given context by the places we live in and the communities we inhabit. Space is one of the principal media through which structure and form are expressed, and spatial organisation produce complex geometries of relationships and structure. Information spaces also induce spatial geometries on the environment they are embedded in. An example of this is the Internet which can be viewed as a spatial network of nodes corresponding to computers, routers, etc, at fixed locations, and edges, corresponding to direct physical or virtual connections. This network structure can be expanded to include the graphical connections associated with the information available through the Internet. In this expanded network structure, the World Wide Web (Web), the nodes represent Internet pages and the edges are associated with hyperlinks connecting information contained in different pages. The power of the Web stems from the linking it makes possible. Several techniques are in use to discern the spatial geometry of the Web. These are broadly classified according to the attributes they measure [1], and are used to discern the Web's topology and its other invariants enabling, inter alia, better browsing and searching experiences.

One of the aims of the Semantic Web is to add more machine-readable semantics to web information via annotations written in a language called the Resource Description Framework (RDF). Although initiatives such as the Linking Open

Data community project[1] are making available vast amounts of interlinked RDF data(e.g. as of 2009 there were 6.7 billion RDF triples and over 140 million links between datasets available [2]), it would be interesting to find out the interlinkage patterns between these datasets and the kinds of spatial geometries they induce. This will help a consumer of linked data or a user of an application that uses linked data to answer questions such as “Where are am I”?, “How did I get here”?, “What words did I use”?, “What data linkages did I traverse to get here”?, “What else are around me that I may find useful”? Deciphering the spatial and topological properties of the linked data web graph will help to answer some of the aforementioned questions as well as help in building effective linked data consumption and federated querying strategies.

Contributions. In this paper,

1. we provide a formal model of the Linking Open Data (LoD) cloud.
2. we carry out and provide a set of network analytic measures for the LoD.
3. the network analytic measures we provide values for are: (a) nodal (out/in)degrees, mean nodal degree, density, and diameter.

2 Data Models for the LoD Cloud

In this section, we present RDF data models and use these as the bases for the model of the LoD cloud. We present a lattice of semantic relationships in the LoD cloud which we use later in the paper as the basis for our network analytic measures. We also introduce the network analytic measures used later in the paper.

2.1 RDF Data Model

In the RDF model, the universe to be modelled is a set of *resources*, identified by Universal Resource Identifiers (URIs), each of which is described in terms of their properties and property values. Descriptions of resources are *statements* in the *subject-predicate-object* triple structure, where subject, predicate, and object are resources. Both subject and object can be anonymous objects, known as *blank nodes*, and object can be strings.

Definition 1. An RDF triple $T =< (R \cup B), U, (R \cup L \cup B) >$ where R is a set of resources, U is a set of URI references, B is a set of anonymous objects (i.e. blank nodes), and L is a set of RDF literals.

Definition 2. An RDF model M is a set of such triples, and is defined as $M \subset (R \cup B) \times U \times (R \cup L \cup B)$.

[1] http://linkeddata.org

Definition 3. An RDF graph $G =< V, E >$ is a labelled directed graph, where $E = V \times P \times V$ is the set of edges of G, V is a subject or an object in G, and P is the set of predicates connecting sets of nodes (V).

2.2 Lattice of Semantic Relationships

The problem of evaluating semantic relatedness using network, graph-like, or controlled vocabularies' representations has a long history in computer science. It is reflected in the work of Quillian, 1968 [3] and of Collins and Loftus, 1975 [4]. In controlled vocabularies, three main types of semantic relationships can be observed. These are:

1. *Hierarchy*: used to state different levels of super-ordination and sub-ordination. For example, car "is-a" kind of vehicle. Example predicates from the Semantic Web include *rdfs:subClassOf,rdfs:subPropertyOf*. About 80% of all semantic relationships are "is-a" relationships [5].
2. *Association*: used to state associations between terms that are neither equivalent nor hierarchical, but the terms are semantically or conceptually associated to such an extent that the link between them should be made explicit in the controlled vocabulary, on the grounds that it may suggest additional terms for use in indexing or retrieval. An example predicate in the Semantic Web is *owl:equivalentClass* which states that two classes have the same class extensions, but are not (necessarily) the same concepts, i.e. they are equivalent extensionally but not intensionally. Two examples: cell/cytology, and artist/musician.
3. *Equivalency*: this is used to state that a concept can be expressed by two or more terms. Examples are synonyms in common words, e.g. cats/felines and freedom/liberty. The predicate used in the Semantic Web is *owl:sameAs*.

Although *owl:sameAs* should be used to represent (strict) identity, it has been shown, in practical Linked Data usage, it is generally misused in this respect. Halpin and Hayes[6] showed four distinct uses of *owl:sameAs* in addition to its recommended usage of strict identity. These include being used as "Same Thing As But Different Context" and "Very Similar To", etc. Vatant[2] suggested, from observations of *owl:sameAs* "in the wild", it is not symmetric.

Definition 4. If H is the set of Hierarchy (uri) predicates, A the set of Association (uri) predicates, and Q is the set of Equivalent (uri) predicates, we define P as the set of all these predicates in E, i.e. $P = H \cup A \cup Q$. P, then, is a preorder on E, such that $E =< P, \sqsubseteq >$ and $Q \sqsubseteq A \sqsubseteq H$. Applying this lattice to semantic relationships in controlled vocabularies gives us the diagram in Figure 1. This figure shows there is a degree of (pre)order and importance of semantic relationships in ontologies.

2.3 Linked Data Cloud Model and Description of Linked Datasets

Figure 2 shows the LoD cloud of the data sets that have been published so far.

[2] http://blog.hubjects.com/2007/07/using-owlsameas-in-linked-data.html

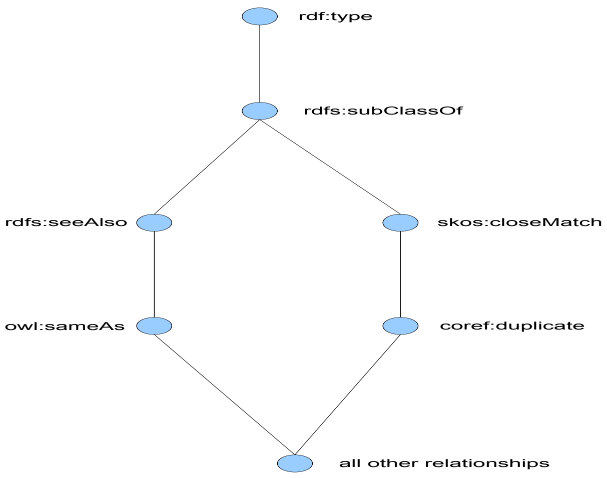

Fig. 1. Lattice map of some Semantic Web vocabulary predicates.

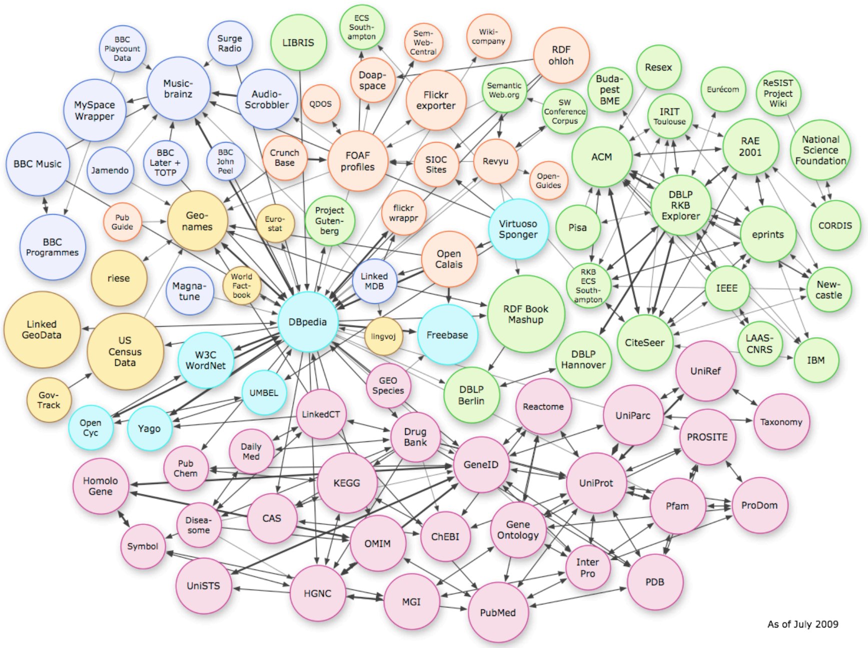

Fig. 2. The Linking Open Data cloud.

Definition 5. An LoD cloud, $\boldsymbol{L}$, is a set of nodes, $\boldsymbol{N}$, linked together by a set of edges, $\boldsymbol{E}$, and is defined as $\boldsymbol{L} = < \boldsymbol{N}, \boldsymbol{E} >$, where $\boldsymbol{E} = \{p \in P : a \in URI(N_i),\, b \in URI(N_j) \wedge a\, p\, b \in N_i \wedge N_i \neq N_j\}$.

2.4 Network Analysis

The network analysis of structures emphasises structural relations as its key orienting principle, where social structure consists of "regularities in the patterns of relations among concrete entities" [9]. Entities may be individual persons, small groups, organisations, etc. There are two fundamental components of network analysis. The first component consists of means for detecting network structures, such as cliques and structurally equivalent positions. The second component involves characterising nodes in the network into various structural terms, such as centrality, prestige, connectedness, etc. Several network analytic metrics are used to decipher structural relations. We mention a few metrics here (which we used later in our analyses).

A **path** from v_1 to v_n is a collection of distinct points, $v_1, v_2, \ldots, v_n$ together with the lines $v_1\, v_2, v_2\, v_3, \ldots, v_{n-1}\, v_n$ considered in the following order: $v_1, v_1\, v_2, v_2, v_2\, v_3, v_3, \ldots, v_{n-1}\, v_n, v_n$. The length of a path is its **path distance**. This notion of path distance is very important in network analysis. In transportation, Internet communication, or the spread of news and diseases, it is often important whether something flowing through a network has to travel just a few hops or many. Several paths may exist between two nodes, but the shortest path between them is the **geodesic**. Geodesic distance is a measure of the nodes' closeness in a network. This is useful for network traversal and network optimisation.

Two nodes are said to be reachable if at least one path of any length exists between them. A graph is connected if paths exist between every pair of nodes, but is disconnected if at least one pair has no path between them. A node that has no lines connecting it to any other node is an **isolate**. There are three types of graph-connectedness:

1. Strongly connected: every pair of nodes is connected by directed paths in both directions (i.e. from v_a to v_b and from v_b to v_a),
2. Unilaterally connected: all pairs are linked by a path in one direction but not in the other direction,
3. Weakly connected: all pairs are joined by lines disregarding their direction.

The **degree** of a point is the integer count or number of other nodes with which a given node has direct contact. The **outdegree** of node v_i is the number (or proportion) of relations from that node to all others. The **indegree** of node n_j is the number (or proportion) of relations received by node n_j from all others. The **nodal degree**, $d(v_i)$, is the total number of relations of the ith node where degree refers to number of lines. The **mean** nodal degree of the network is obtained by summing the nodes' nodal degrees and dividing by the number of nodes. Mean nodal degree is useful to know "who knows who". The **prestige** of a node in a network is the extent to which a node in a network “receives” or “serves as the object” of relations sent by others in the network. The sender-receiver or source-target distinction strongly emphasises inequalities in control over resources. The **density** of a network is the total number of all relations for all nodes in the network divided by the number of all possible relations.

3 Network Analyses of the LoD - Experimental Setup

In any network analysis experiment, there is the question of boundary specification, i.e. where does one set the limits of the network. We included datasets that had queryable SPARQL[3] endpoints and/or RDF data dumps. Our experiment can be divided into these phases:

1. Manual traversal of each node in the LoD, to find out which nodes of the LoD can be queried, if they have reachable SPARQL endpoints, and how we can find out about the structures of the RDF instances in these nodes.
2. Ascertaining Instances of Lattice of Semantic Relationships from query-able nodes, RDF dumps, and voiD files.

3.1 Traversal of each LoD node

This was a manual traversal[4] where we visited the URL of each node mentioned at the LoD cloud[5]. This manual traversal turned up a few observations. We noticed the following:

1. SIOC, FOAF, and DOAP are not data stores or knowledge bases. These are vocabularies used to represent information and knowledge that are inserted (or asserted) into knowledge bases. These vocabularies are scattered throughout the Semantic Web, they have no specific knowledge base and we therefore decided to exclude them from our analyses.
2. Wikicompany. This is a placeholder for DBpedia[6]. It was found to be empty, and is therefore eliminated from our analyses.
3. PubGuide. This was down for “maintenance” during the times of our experiments, so we eliminated this from our analyses.
4. Flick exporter. This was no longer there at the times of our experiments, so we eliminated this from our analyses.
5. Eurostat: This was down during the times of our experiments, so we eliminated this from our analyses.
6. Rdfohloh had no (RDF) dataset dump nor SPARQL endpoint to exploit, and is therefore eliminated from our analyses.

[3] http://www.w3.org/TR/rdf-sparql-query/

[4] We carried out our experiments between the months of May to July 2010.

[5] http://richard.cyganiak.de/2007/10/lod/

[6] http://dbpedia.org

7. http://www.surgeradio.co.uk/ : No (RDF) dataset nor SPARQL endpoint was found. Email sent to the email address given on the site. This bounced back. So, we eliminated this from our analyses.
8. http://openguides.org/: No (RDF) dataset dump nor SPARQL endpoint was found. So, we eliminated this from our analyses.
9. http://qdos.com/: No (RDF) dataset dump nor SPARQL endpoint was found. So, we eliminated this from our analyses.

Other Observations: We noticed that some of the nodes in the cloud are sub-domains of other larger nodes. These are listed in table 1.

Topics	**Larger Node**	**Datasets**
Biology/Genetics	BIO2RDF	PubMed, MGI, PDB, HGNC, Unists Symbol (Gave an error “...server not found ...”) Homologene, Taxonomy, Prodom, Interpro, Omim, Cas (Gave an error “...server not found ...”) Pubchem, Uniref, Uniparc, Prosite, Pfam, GO Chebi, Reactome, GeneId
Entertainment/Music	DBTUNE	Jamendo, MusicBrainz, Magnatune virtuoso.dbtune.org (Gave ”Service Temporarily Unavailable” error) BBC Playcount Data, MySpace Wrapper, BBC John Peel
Researchers/Publications	DBLP RKB Explorer	Budapest, Resex, Eurecom, Resist Project Wiki National Science Foundation, CORDIS, eprints, Newcastle, IEEE, IBM, RKB ECS Southampton, Pisa, ACM IRIT Toulouse, RAE 2001, Citeseer, LAAS-CNRS
Medical/Research/Publications	Wiwiss.fu-berlin.de	Diseasome, DailyMed, DBLP, Drugbank Factbook, Gutendata

Table 1. Larger Node - Subnode pairing.

3.2 Ascertaining Instances of Lattice of Semantic Relationships from query-able nodes, RDF dumps, and voiD files

From our initial traversal of all the LoD nodes, we noticed some nodes have accessible SPAQRL endpoints, while some provide RDF data dumps of their data. We obtained voiD files for DBpedia and RKBExplorer, which we used to analyse and explore these two nodes. voiD is an RDF based schema used to describe the content and the interlinking between datasets.

Unexplored Nodes. We did not explore OpenCalais as we found no sparql endpoint nor RDF data dumps. On the OpenCalais website, we found a statement which read “At present, there is no way to query the Calais Linked Data

System to see if an entry for a particular entity exists."[7] It also said[8], "The assets Calais currently links to are: DBpedia, Wikipedia, Freebase, Reuters.com, GeoNames, Shopping.com, IMDB, LinkedMDB." It is difficult to establish the veracity of this statement, and difficult to establish the relationship types. Freebase does not have a SPARQL endpoint neither does it have RDF data dumps (although it has data dumps that are not RDF). It was difficult to find details of outgoing links from Freebase to other LoD nodes. We decided to use the incoming links to Freebase and OpenCalais in our analyses.

4 Analysis/Evaluation

Section 2.2 gave us a lattice of semantic relationships (*rdf:type, rdfs:subClassOf, rdfs:seeAlso, owl:sameAs, skos:closeMatch, coref:duplicate*). We used this lattice for our queries either on the nodes' sparql endpoints or with the provided RDF dumps. We kept note of the foreign links found on each node. For the voiD files, we noted the linkset relationships as found in the voiD files.

From our experiments, we get the linkages and the linkage types shown in figure 3. We notice that although the LoD is connected, it is **weakly connected**. If we disregard directionality, DBLP Hannover is **structurally related** to DBLP Berlin, as both have similar relations to equivalent other network nodes. The notion of structurally equivalent datasets will be important in the coming months as new datasets join the cloud and the ability to surface datasets of similar functionality will be important. Although there is no isolate in the LoD, we observe the most **isolated** nodes are GeneOntology, Revyu, Riese, GovTrack.us, and LinkedLifeData. The reason for not having a competely isolated node is a prerequisite for inclusion is that a dataset must be linked to at least one other dataset. If we ignore directionality, we notice **transitivity** in the LoD (i.e. if A connects to B and B connects to A, then A connects to B). The significance of this is that transitivity makes it easier to query and traverse the LoD, but this should not be used as a basis for inference because the predicates used for traversal may not have good closure properties. We notice that most of the links (about 90%) are intralinks within a dataset. We observe very few inter-dataset linkages. This is similar to the Web where over 75% of hyperlinks connect pages on the same host [8].

We notice that the most **prestigious** node is DBpedia, as it has the highest **indegree** number of 18 (not taking account of directionality). This shows the importance and centrality of DBpedia in the LoD, and it also shows that a node, taken at random, is likely to have a DBpedia linkage. This is not unexpected as DBpedia is structured information extracted from Wikipedia. As Wikipedia covers a broad range of concepts, DBpedia, as its linked data equivalent, is expected to have a similarly broad range of concepts and is expected that other datasets will link to it in order to use its concepts. We observe Virtuoso Sponger

[7] from http://www.opencalais.com/documentation/calais-linked-data/linkfaq

[8] http://www.opencalais.com/documentation/linked-data-entities

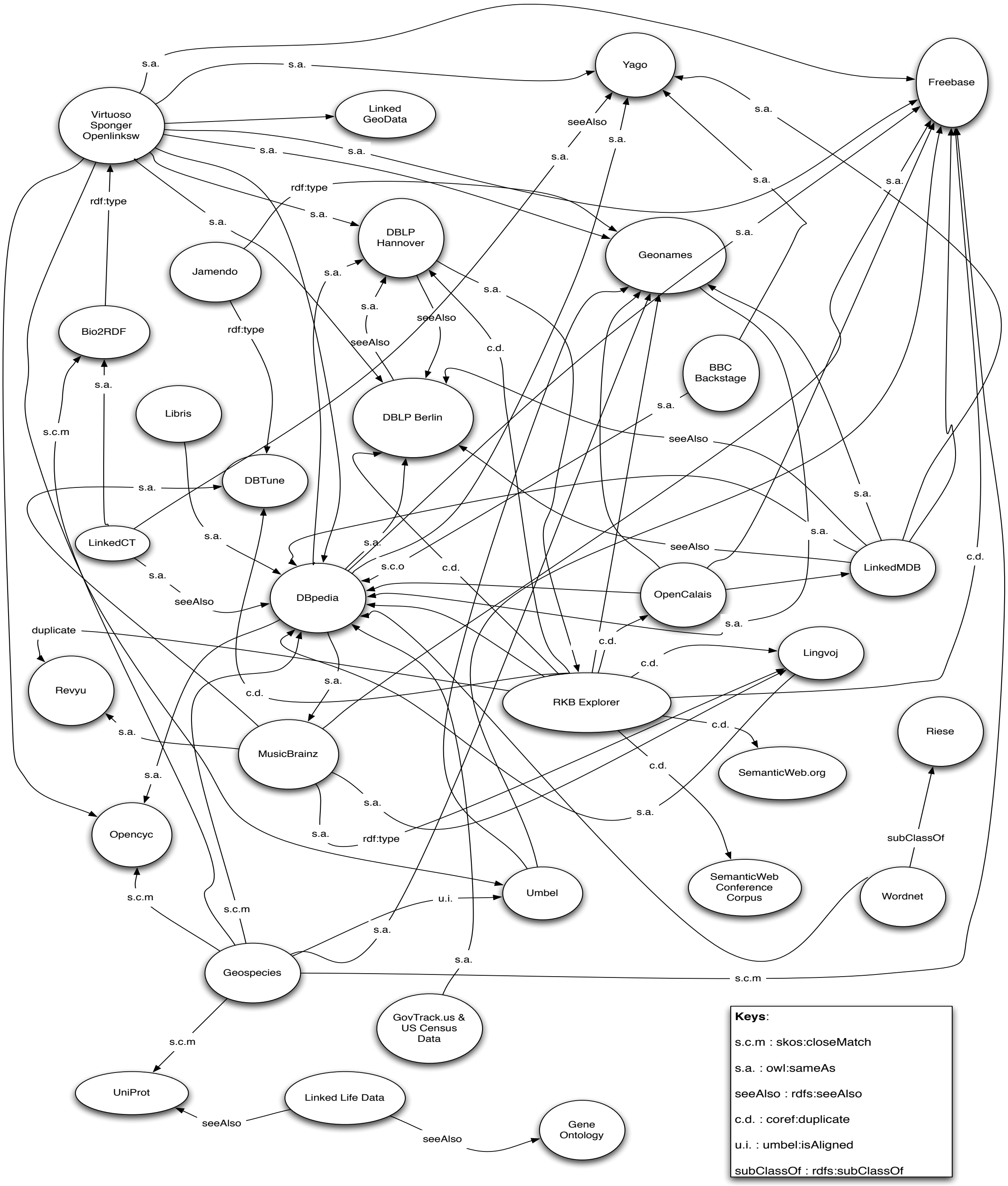

Fig. 3. Some linkages and the linkage types in the LoD

has the highest **outdegree** number of 10, so therefore it is the most **felicitate**[9] of the LoD. Virtuoso Sponger generates RDF Linked Data from a variety of data sources, and inserts (asserts) these RDF data into its knowledge bases, and therefore is expected to have more outgoing links than other datasets. A possible consequence of this is that a new dataset coming into the LoD, using the principle of least effort[10], may decide to link to Virtuoso Sponger as it (i.e. Virtuoso) is more likely to be linked to a high number of datasets in the cloud.

Table 2 shows the approximate number of linkages and their types in some of the datasets. For some datasates, we were only able to use their voiD files, e.g. RKB Explorer. Table 3 shows the nodal degree of the LoD elements. The **mean nodal degree** is ≈ 4, while the **density** is ≈ 0.13. This low density lends credence to the weak connectivity of the LoD (density normally takes values between 0 [totally disconnected] and 1 [strongly connected]). The **geodesic** of the LoD is found to be 4. This is small compared to the geodesic of the Web, which was found to be between 19 and 21[10]. While we expect this value to increase as the cloud becomes larger, a small geodesic value is a feature of a *small-world* network, as seen in early analysis of Internet data [13].

Node	**rdf:type** links	**rdfs:seeAlso**	**rdfs:subClassOf**	**owl:sameAs**
BIO2RDF	5372	17080	12285	
Jamendo				1879
Libris				1000
DBLP Berlin		490		500
LinkedCT		20301		10888
LinkedMDB				12207
LinkedLifeData	63	120		

Table 2. Approximate linkage numbers and type.

5 Related Work

Topological measures of the Web of Data have not been done extensively. The LoD cloud[11] was generated from statistics[12] hand-compiled from data from project home pages, asking dataset owners, etc. A number of arbitrary decisions were made in the process[13], e.g. how many links should one count to warrant a linkage to a different dataset, should a huge project such as Bio2RDF be represented as one big bubble or as many little ones, etc. Hausenblas et. al. [11] asked the question "What is the size of the Semantic Web?", but provided no

[9] http://www.drbilllong.com/SpellersDiary2/Felicity.html

[10] http://en.wikipedia.org/wiki/Principle_of_least_effort

[11] http://richard.cyganiak.de/2007/10/lod/

[12] http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics

[13] http://lists.w3.org/Archives/Public/public-lod/2009Aug/0069.html

Node	Nodal degree	Node	Nodal degree
GovTrack/USCensus Data	1	Gene Ontology	1
Riese	1	Libris	1
LinkedGeoData	1	SemanticWeb.org	1
SemanticWeb Conference Corpus	1	Jamendo	2
BBC Backstage	2	Revyu	2
LinkedLifeData	2	UniProt	2
WordNet	2	Bio2RDF	3
Lingvoj	3	LinkedCT	3
DBTune	3	UMBEL	4
Opencyc	4	MusicBrainz	4
OpenCalais	4	Yago	5
Freebase	6	DBLP Hannover	6
DBLP Berlin	6	LinkedMDB	7
Geospecies	7	Geonames	8
Virtuoso Sponger	11	RKB Explorer	12
DBPedia	18		

Table 3. Node - Nodal Degree.

specific value. Their provision of no value for the size may be due to the high variability of the LoD cloud at the time they asked the question. Gil and Garcia [12] used statistical mechanics methods to measure the size of the Semantic Web. They used an RDF crawler to measure the geodesic of 1.4 million triples for 282 ontologies of the DAML Ontology Library[14]. Their geodesic value was given as 4.37. This is a value that is comparable to our, and further validates the small-world property of the Semantic Web.

6 Conclusion and Future Work

This paper provided a characterisation of the Web of Data. We investigated the Linked Open Data (LoD) cloud to study its topology. We provided a formal model of the LoD, and observed that it induces a weakly connected graph and has a fundamental feature of a *small-world* due to its small geodesic value. From our investigation of the LoD, we provided values for a number of network analytic measures, including its mean nodal degree, its geodesic value, and the node that is of the highest prestige. The most felicitate node was Virtuoso Sponger while the most prestigious node was DBpedia. As new datasets are added to the cloud, it will be interesting to find out how they affect the topological properties. Would a new prestige node/dataset emerge or would the model of preferential attachment of Barabasi and Albert [14] favour DBpedia and enhance its prestige? It would also be interesting to find out if there is an asymptotic value from which these topological properties diverge.

[14] http://www.daml.org/ontologies

For future work, we intend to use this topological relationship to capture a "Relationship MetaLayer" atop the LoD. Topological relationships have played an important role in query languages [15], we intend to apply this LoD semantics towards effective query processing.

7 Acknowledgements

This work was supported by the EnAKTing project, funded by EPSRC project number EP/G008493/1. We thank Ross Horne for his contributions.

References

1. D. Dhyani and N. W. Keong: *A Survey of Web Metrics.* ACM Computing Surveys, Volume 34 , Issue 4, pp. 469 - 503, 2002.
2. C. Bizer: *The Emerging Web of Linked Data.* In IEEE Intelligent Systems, Volume 24, no. 5, pp. 87-92, September/October, 2009.
3. M. Ross Quillian: *Semantic Memory.* In M. Minksy, *Semantic Information Processing*, 1968.
4. A. Collins and E. Loftus: *A spreading activation theory of semantic processing.* In Psychological Review, Volume 82, pp. 407-428, 1975.
5. Sh. Shariatmadari, A. Mamat, H. Ibrahim, N. Mustapha: *Ranking Semantic Similarity in Semantic Web.* In Poster and Demo Session Proceedings of the 7th International Semantic Web Conference, (ISWC 2008), 2008.
6. H. Halpin and P. J. Hayes: *When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web.* In Linked Data on the Web (LDOW2010).
7. K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao: *Describing Linked Datasets - On the Design and Usage of voiD, the "Vocabulary of Interlinked Datasets".* In Linked Data on the Web (LDOW 09), 2009.
8. K. Bharat, B-W Chang, M. Henzinger, M. Ruhl: *Who Links to Whom: Mining Linkage between Web Sites.* Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 51-58, 2001.
9. D.R. White, A. Boorman, and R. L. Brieger: *Social Structure from multiple networks 1.* American Journal of Sociology 81, 1976.
10. R. Albert, H. Jeong, and A-L Barabasi: *Internet: Diameter of the World-Wide Web.* Nature 401, 130-131 (9 September 1999).
11. M. Hausenblas, W. Halb, Y. Raimond, and T. Heath: *What is the Size of the Semantic Web?.* In I-Semantics 2008: International Conference on Semantic Systems, 2008.
12. R. Gil and R. Garcia: *Measuring the Semantic Web.* In SIGSEMIS Bulletin Vol 1, Issue 2, pp. 69-72, 2004.
13. M. Faloutsos, P. Faloutsos, and C. Faloutsos: *On power-law relationship of the Internet topology.* In Computer Communications Review, Volume 29, pp. 251-263, 1999.
14. A-L. Barabasi and R. Albert: *Emergence of scaling in random networks.* In Science Volume 286, pp. 509-511, 1999.
15. E. Clementini, J. Sharma, amd M. J. Egenhofer: *Modelling topological spatial relations: Strategies for query processing.* Computers & Graphics, vol. 18, issue 6, 1994.