Object Link Structure in the Semantic Web

Weiyi Ge[1], Jianfeng Chen[1], Wei Hu[2], and Yuzhong Qu[2]

[1] School of Computer Science and Engineering, Southeast University, China
[2] State Key Laboratory for Novel Software Technology, Nanjing University, China
wyge@seu.edu.cn, jferic@seu.edu.cn, whu@nju.edu.cn, yzqu@nju.edu.cn

Abstract. Lots of RDF data have been published in the Semantic Web. The RDF data model, together with the decentralized linkage nature of the Semantic Web, brings object link structure to the worldwide scope. Object links are critical to the Semantic Web and the macroscopic properties of object links are helpful for better understanding the current Data Web. In this paper, we propose a notion of object link graph (OLG) in the Semantic Web, and analyze the complex network structure of an OLG constructed from the latest dataset (FC09) collected by the Falcons search engine. We find that the OLG has the scale-free nature and the approximate effective diameter of the graph is small compared to its scale, which are also consistent with the experimental result based on our last year's dataset (FC08). The amount of RDF documents and objects by Falcons both doubled during the past year, but the object link graph remains the same density while the diameter is getting shrinking. We also repeat the complex network analysis on the two largest domain-specific subsets of FC09, namely Bio2RDF(FC09) and DBpedia(FC09). The results show that both Bio2RDF(FC09) and DBpedia(FC09) have low density in object links, which contribute to the low density of object links in FC09.

1 Introduction

In recent years, more and more RDF data have been published in the Web, and most of them are created to describe objects by using shared classes and properties. From Aug. 2008 to Sept. 2009, the number of RDF documents collected by the Falcons search engine [8] increases from 11.7M to 21.6M, with the number of RDF triples from 600 million to 2.9 billion, as well as the numbers of objects, classes and properties from 73.8M, 2.2M ,203K to 171.4M, 2.8M, 264K respectively.[1] As pointed out in [15], the Web is being extended with more and more RDF data sources and links between objects, even across data sources. The RDF data model, together with the decentralized linkage nature of the Semantic Web, brings object link structure to the worldwide scope, where objects are identified by URIs, and links are attributed to relational properties among objects.

The hypertext Web is considered to be a directed graph whose vertices correspond to Web pages and arcs correspond to hyperlinks between the pages, so

[1] http://ws.nju.edu.cn/falcons/statistics.jsp

L. Aroyo et al. (Eds.): ESWC 2010, Part II, LNCS 6089, pp. 257–271, 2010.

the page link graph in the hypertext Web is formed. We believe that the object link structure is important to the Semantic Web, as the Web page link structure to the hypertext Web.

Complex network analysis has been extensively performed on the page link graph to reveal the macroscopic properties of the hypertext Web [1,2,3,6,12]. Recently, graph analysis techniques have also been applied to the schema level of the Semantic Web, from single ontologies to a set of ontologies, even to the whole Semantic Web [9,13,16,21,24]. However, to the best of our knowledge, the macrostructure of the instance level of the Semantic Web has not yet been well studied. We argue that a simple link structure of object links reveals some useful macroscopic properties, which needs to be studied so that we can better understand the macrostructure of the current Data Web.

In this paper, we propose a notion of object link graph (OLG) in the Semantic Web, and analyze the complex network structure of an OLG constructed from the latest dataset (FC09) collected by Falcons until Sept. 2nd, 2009. We find that the OLG has the scale-free nature and the approximate effective diameter of the graph is small compared to its scale. Then, by comparing this OLG with another one constructed with our last year's dataset (FC08), we confirm our findings. Besides, the amount of RDF documents and objects in Falcons both doubled during the past year, but the object link graph remains the same density and its diameter is getting shrinking, which indicates a good evolution of the Data Web. We also repeat the complex network analysis of OLG on the two largest domain-specific subsets of FC09, namely Bio2RDF(FC09) and DBpedia(FC09). The results show that both of Bio2RDF(FC09) and DBpedia(FC09) have a low density in object links, which contributes to the low density of object links in FC09.

The remainder of this paper is organized as follows. Section 2 gives basic terminology used in this paper. Section 3 provides an overview of datasets used in the experiments and introduces our experimental methodology. Section 4 analyzes the OLG constructed from FC09. Section 5 investigates the evolution of the object link graph in the past year. In Section 6, we extract two domain-specific OLGs and compare their structures with OLG in FC09. Section 7 discusses related work. Section 8 concludes the paper with some observations and possible future work.

2 Terminologies

2.1 Graph

An undirected graph consists of a finite nonempty set of vertices V and a set of edges E. An edge in E is an unordered pair (u, v) representing a connection between two vertices $u \in V$ and $v \in V$.

A connected component of an undirected graph is a subgraph in which any two vertices are connected to each other by paths, and to which no more vertices or edges can be added while preserving its connectivity. The number of vertices in the connected component is called its size.

For each natural number d, let $g(d)$ denote the fraction of connected node pairs whose shortest connecting path has length at most d. The hop-plot for the graph is the set of pairs $(d, g(d))$, which denotes the cumulative distribution of distances between connected node pairs. We extend the hop-plot to a function defined over all positive real numbers by linearly interpolating between the points $(d, g(d))$ and $(d+1, g(d+1))$ for each d, and we define the effective diameter of the graph to be the value of d at which this function achieves the value 0.9 [19].

A random variable x is distributed according to a power law when its probability density function $p(x)$ is given by $p(x) = Ax^{-\gamma}$, where A and γ are positive constants, and γ is called the power law exponent. A power law distribution plotted on a log-log scale is a line. A graph whose degree distribution follows a power law is scale-free.

2.2 Objects in the Semantic Web

An entity is a named resource identified by a URI in RDF data. An entity e is regarded as a class (or a property) in an RDF document if the RDF graph encoded in the document entails the RDF triple $\langle e$, `rdf:type`, `rdfs:Class`$\rangle$ (or $\langle e$, `rdf:type`, `rdf:Property`$\rangle$), and it is regarded as an object if it is neither a class nor a property. In accordance with [9], we require the disjointedness of classes, properties, and objects, similar to OWL DL.

There may be more than one RDF documents in the Semantic Web that describe the same resource but give inconsistent description. For example, it is possible that a URI is stated to identify an object in one RDF document but to identify a class in another document. Inspired by [9], we developed the following heuristics to resolve the inconsistency. Firstly, we determine the identity of a URI by considering only its dereference document. If such document is not available, we will consider that a URI identifies an object only if no documents states that it identifies a class or property.

It is noteworthy that we are only interested in named resources but ignore blank nodes because a blank node cannot be directly referred outside the RDF graph it is defined by. Thus, cross-document links never happen to blank nodes so that they are not considered in the following Web-scale analysis. However, blank nodes may indirectly contribute to the Web-scale object link structure, which will be discussed in the next subsection.

2.3 Object Link Graph

An object link graph, denoted by $(\mathbb{O}, \mathbb{L})$, is an undirected graph, where $\mathbb{O}$ is the vertex set, each is identified by a unique URI to represent an object; $\mathbb{L}$ is the edge set, and each edge (u, v) exists iff there is a sequence of k triples $\{\langle a_i, p_i, b_i\rangle | 1 \leq i \leq k\}$, where $(a_1, b_k) \in \{(u, v), (v, u)\}$ and $b_i = a_{i+1}$, b_i are all blank nodes for $1 \leq i \leq k-1$. Here, we do not simply assume the directionality of links between objects. According to this definition, blank nodes are not included in an object link graph, but blank nodes may still contribute to establishing links between objects.

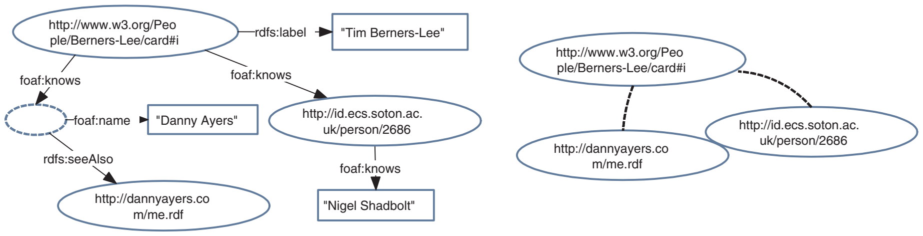

http://www.w3.org/People/Berners-Lee/card#i
rdfs:label
"Tim Berners-Lee"
foaf:knows
foaf:knows
foaf:name
"Danny Ayers"
http://id.ecs.soton.ac.uk/person/2686
rdfs:seeAlso
foaf:knows
http://dannyayers.com/me.rdf
"Nigel Shadbolt"
http://www.w3.org/People/Berners-Lee/card#i
http://id.ecs.soton.ac.uk/person/2686
http://dannyayers.com/me.rdf

Fig. 1. An RDF graph (a) and its corresponding object link graph (b)

Fig. 1(a) shows a fragment of an RDF graph, which is derived from an RDF document identified by `http://www.w3.org/People/Berners-Lee/card`. From this RDF graph, we obtain an object link graph, as shown in Fig. 1(b). In particular, the link from `http://www.w3.org/People/Berners-Lee/card#i` to `http://dannyayers.com/me.rdf` is attributed to the fact that there is a simple path of length 2 via a blank node.

3 Datasets and Experimental Methodology

All the experimental results presented in this paper are obtained by analyzing datasets collected by the Falcons search engine. This section firstly gives an overview of the Falcons crawler. Then, datasets used in this paper are introduced. At last, experimental procedure is presented.

3.1 Crawler

To ensure the coverage of the Falcons crawler, we feed it a set of seed URIs of RDF documents from three sources. Firstly, we extract some keywords from the Open Directory Project,[2] and randomly combine them as queries in Swoogle[3] and Google (for “filetype:rdf” and “filetype:owl”) to retrieve URIs of potential RDF documents. Secondly, as many personal RDF data are stored on several online repositories such as `pingthesemanticweb.com`, URIs of RDF documents from these repositories are added to the seed set. Thirdly, as the largest data source in the current Semantic Web, several entry-point URIs of the datasets published in the Linking Open Data project are manually submitted.

A parallel crawler is implemented to dereference URIs with content negotiation (ACCEPT application/rdf+xml) and download RDF/XML documents. Besides, this crawler follows the robots.txt protocol with the HTTP header field User-Agent setting to “Falconsbot”. We do not use any filter rules or domain limits, as we want to get a relatively complete sample of the current Data Web.

[2] `http://www.dmoz.org/`

[3] A representative Semantic Web search engine, `http://swoogle.umbc.edu/`

3.2 Datasets

We exported two snapshots of datasets from the Falcons crawler in August 26th, 2008 and September 2nd, 2009. The 2008's dataset contains 11,719,608 RDF documents, and the 2009's dataset contains 21,639,337 RDF documents.

```
<foaf:Person rdf:nodeID=“me”>
  <foaf:nick>Carlita</foaf:nick>
  <foaf:knows>
    <foaf:Person>
      <foaf:nick>Gabriela</foaf:nick>
      <rdfs:seeAlso rdf:resource=“http://api.hi5.com/rest/profile/foaf/212231607”/>
    </foaf:Person>
  </foaf:knows>
</foaf:Person>
```

Fig. 2. A fragment taken from an RDF document in hi5.com

After inspecting the top-50 domains[4] with most RDF documents, we find some social networking sites have the similar publishing style, which use blank nodes to identify objects instead of URIs. Fig. 2 shows a fragment taken from an RDF document in `hi5.com`. In the fragment, a person with nickname “Carlita” (blank node) knows another person “Gabriela” (blank node), and Gabriela has an FOAF document `http://api.hi5.com/rest/profile/foaf/212231607`. Here, blank nodes are used to identify persons. It is stressed that this publishing way seems incompatible with the Linked Data principles [5], since it does not use URIs to identify things and cannot interlink objects across data sources. Most of RDF documents from `hi5.com`, `mybloglog.com`, `buzznet.com`, `liveinternet.ru`, `deadjournal.com` and `rambler.ru` have the similar publishing style. So, we refine the datasets by excluding those RDF documents that come from the above six domains, and the refined datasets are called **FC08 (Falcons Crawl 2008)** and **FC09 (Falcons Crawl 2009)** respectively.

We find that FC08 has 11,286,186 RDF documents coming from 10,216 domains and FC09 has 18,646,011 RDF documents from 21,171 domains. That is, the FC09 dataset is doubled in the quantity of documents and the diversity of domains as compared with FC08. The top-10 domains w.r.t. the number of RDF documents are listed in Table 1.

From FC08, we identify 64,974,423 objects by using the heuristics described in Section 2.2. Among all the objects identified, 63,795,076 ones (98.18%) are identified by the HTTP URIs, and they are hosted by 621,619 domains. From FC09, we identify 110,507,074 objects, and 108,842,826 ones (98.49%) are identified by the HTTP URIs hosted by 698,753 domains. That is, the number of

[4] The domain name of the URL is the substring of the URL's hostname, without subdomain names. For example, `fu-berlin.de` is the domain name of `http://www4.wiwiss.fu-berlin.de/dblp/terms.rdf`

Table 1. Top-10 domains w.r.t. the number of RDF documents

(a) FC08

Domain	#documents
`bio2rdf.org`	6,636,748
`dbpedia.org`	2,577,748
`opiumfield.com`	415,534
`geonames.org`	359,684
`w3.org`	211,388
`l3s.de`	156,786
`fu-berlin.de`	129,187
`bibsonomy.org`	113,650
`rkbexplorer.com`	110,524
`uniprot.org`	98,159

(b) FC09

Domain	#documents
`bio2rdf.org`	7,685,644
`dbpedia.org`	3,712,453
`geonames.org`	1,497,089
`opiumfield.com`	785,223
`l3s.de`	719,138
`dbtune.org`	577,634
`fu-berlin.de`	564,848
`openlinksw.com`	498,047
`bibsonomy.org`	398,665
`w3.org`	392,516

Table 2. Top-10 domains w.r.t. the number of objects

(a) FC08

Domain	#objects
`bio2rdf.org`	28,276,823
`dbpedia.org`	6,671,120
`wikipedia.org`	3,955,286
`flickr.com`	2,501,768
`fu-berlin.de`	2,282,677
`uniprot.org`	1,931,325
`l3s.de`	1,842,994
`uni-trier.de`	1,464,458
`musicbrainz.org`	1,332,336
`opiumfield.com`	1,299,949

(b) FC09

Domain	#objects
`bio2rdf.org`	33,667,558
`dbpedia.org`	7,645,474
`opiumfield.com`	6,560,575
`flickr.com`	6,156,454
`last.fm`	5,639,584
`l3s.de`	5,404,294
`wikipedia.org`	4,195,842
`fu-berlin.de`	3,972,447
`dbtune.org`	3,659,400
`geonames.org`	2,803,359

objects is also two times more than the one of last year, and the distribution of these objects is more diverse. The top-10 domains w.r.t. the number of objects are listed in Table 2.

From Table 1(b), we find that `bio2rdf.org` and `dbpedia.org` have most RDF documents (41.22% and 19.91% respectively). Besides, most objects are also distributed in these two domains (30.47% and 6.92%), see Table 2(b). So, we export all RDF documents in `bio2rdf.org` and `dbpedia.org` to form two domain-specific datasets, namely **Bio2RDF(FC09)** and **DBpedia(FC09)**.

In Bio2RDF(FC09), we identify 36,036,254 objects derived from 7,685,644 documents. These objects are distributed in 1,720 domains. Besides, from 3,712,453 documents in DBpedia(FC09), we identify 17,414,639 objects, which are distributed in 505,132 domains. In a sense, this indicates that DBpedia is a linking-hub for interconnecting objects from various data sources.

3.3 Experimental Methodology

Three experiments are designed to explore object link structures in the Semantic Web.

Firstly, to understand the macroscopic characteristics of the object link structure in the current Semantic Web, we analyze the object link graph in FC09. We compute the average degree of this graph, which reflects the density of graph in some sense. Then, we analyze the distribution of isolated vertices and investigate reasons why there are so many isolated vertices. Degree distribution of the graph is depicted to reveal whether the graph has the scale-free natural. Connected components of the graph are computed to reveal whether objects in the Semantic Web are well interlinked. Effective diameter is approximately calculated to seek to understand within how many certain numbers of hops that most connected object pairs can be interlinked.

Secondly, in order to confirm our findings and to find the evolution of OLG's structure, we repeat the complex network analysis on FC08. We compare the network characteristics of two objects link graphs.

In the third experiment, we analyze two domain-specific OLGs constructed from Bio2RDF(FC09) and DBpedia(FC09) in the same way, and compare the experimental results with the OLG in FC09.

4 Object Link Graph in the Current Semantic Web

In this section, we analysis the degree and connectivity of the OLG from FC09.

4.1 Degree

This OLG contains 110,507,074 vertices and 190,201,590 edges, the average degree of the graph is 3.44, and the highest degree is 1,181,411. Broder, et al. [6] report that the web graph constructed from the traditional hypertext Web takes an average in(out)-degree of 7.86, which means its degree as an undirected graph maybe nearly doubled. Compared to the traditional hypertext Web, the small average degree of the OLG indicates a sparse link structure.

By inspecting the degree of each vertex, we find that 4,746,095 (4.29%) objects have no links to others. That is, objects do not link to other objects through a sequence of triples connected by blank nodes (see Section 2.3 for details about constructions of OLG). Most of these isolated objects are distributed in 10 domains, as listed in Table 3. In particularly, the `bio2rdf.org` domain takes 43.6% of all isolated objects found so far. After a close investigation on the RDF documents that mention these isolated objects, we find some typical patterns causing the isolation: a) Objects are not connected to any other objects, but might be connected to classes or properties, e.g. in their type declarations; b) Objects are not connected to any other objects, but connected to literals, e.g. with only literal description about these objects; and c) Objects are connected with blank nodes.

Table 3. Top-10 domains with most isolated vertices in the OLG from FC09

Domain	#isolated vertices
`bio2rdf.org`	2,071,885
`last.fm`	487,137
`yahoofs.com`	271,828
`nbii.gov`	199,979
`friendfeed.com`	160,656
`opencyc.org`	82,059
`umbc.edu`	79,392
`zitgist.com`	76,798
`mpii.de`	57,754
`rossia.org`	55,689

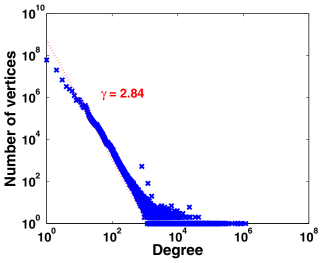


Fig. 3. Degree distribution of the OLG from FC09

The degree distribution is illustrated in Fig. 3. This distribution follows a power law, indicating the scale-free nature of the graph. The power law exponent of degree distributions of the OLG is 2.84, here we use a maximum-likelihood method to fit the exponent [7]. Since the vast majority of vertices in the scale-free network are with small degree, the graph is fault tolerant in the face of random failures; But if a few major high degree vertices (hubs) are removed, it may turn into a set of rather isolated graphs. Besides, anther important characteristic of scale-free network is the clustering coefficient distribution, which decreases as the vertex degree increase. That is, the low-degree vertices belong to very dense sub-graphs and in the meantime hubs connect these sub-graphs together. Since hubs connect sub-graphs each other and the number of hubs is few, we can feed them as seeds in the search engine as a good placement. In-degree and out-degree distributions of subsets of the the traditional Web follow power law with exponents 2.1 and 2.38-2.72, respectively [3,18,6]. OLG's power law exponent is a little larger than ones of the traditional Web.

4.2 Connectivity

Connected Component. Connected component analysis on the OLG shows that the largest CC has 97,391,271 (88.13%) objects. That is, 88.13% of objects in FC09 are reachable to each other by following RDF links. This value is close to the one of the traditional Web (91%) reported in [6], so the connectivity of the OLG is not bad. Except the largest CC and the trivial ones (with only one vertex), there are also 813,975 connected components. Objects in these connected components mainly come from only one single domain (62.12%), and objects in the most one only come from 1,000 domains.

Effective Diameter. Computing the diameter of a large graph is very costly. In the case of the OLG, we cannot compute the exact diameter, instead, we apply the Approximate Neighborhood Function (ANF) approach [20] to estimate the effective diameter for the OLG.[5] The approximate effective diameter of the OLG is 11.53, which is a small one compared to the scale of the graph. Broder, et al. [6] find that the average path length in the traditional Web (when a path exists) is around 6.83 if all edges are considered to be undirected, which is much less than the effective diameter of our OLG. That is, there is an average longer length of the shortest path between connected objects than that of the traditional Web..

5 Structural Evolution of Object Link Graph

As mentioned above, in Section 3.2, the amounts of RDF documents and objects in Falcons Crawl get doubled in the past year. So, we are naturally concern about the structural change brought by the increasing RDF data. In this section, we analyze the object link graph constructed from FC08, and compare it with the OLG from FC09.

Table 4. Top-10 domains with most isolated vertices in the OLG from FC08

Domain	#isolated vertices
`bio2rdf.org`	2,078,204
`li.ru`	677,779
`dbpedia.org`	250,764
`llnwd.net`	91,914
`truesense.net`	91,265
`cyc.com`	43,568
`dbtune.org`	40,253
`mcdonaldbradley.com`	39,072
`rkbexplorer.com`	37,019
`klab.lv`	36,158

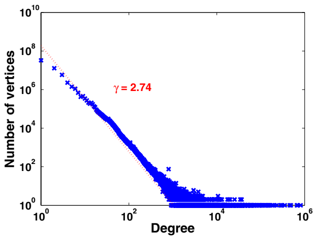


Fig. 4. Degree distribution of the OLG from FC08

5.1 Degree

From 11,286,186 RDF documents in FC08, we identify 64,974,423 objects and construct an object link graph with 64,974,423 vertices and 109,373,275 edges. The average degree of the OLG is 3.37, which is a little smaller than the one (3.44) of OLG from FC09. The slightly increasing average degree indicates that the object link graph become less sparser during the past year.

We find that 3,987,843 (6.14%) isolated vertices are distributed in 16,353 domains. Table 4 lists the top-10 domains containing most isolated objects. Besides, comparing with isolated vertices of the OLG from FC09, we find that some isolated vertices in FC08 disappear in FC09, while some new isolated vertices emerge in FC09. The percentage of isolated vertices declines in the past year (from 6.14% to 4.29%), which shows a trend that objects are getting interlinked.

[5] In our experiment, k is set to 32, which ensures that ANF achieves less than 10% errors.

The degree distribution of the OLG from FC08 is depicted in Fig. 4, which also follows a power law, with the power law exponent 2.74. This exponent is similar to the one of the OLG from FC09 (2.84).

5.2 Connectivity

Connected Component. Connected component analysis on the OLG in FC08 shows that the largest connected component takes 57,122,054 (87.91%) objects and 104,675,519 (95.70%) edges. Except the largest connected component and the trivial ones, there are also 686,071 connected components. Objects in these connected components mainly come from only one single domain (46.72%), and objects in the most one also come from 1,000 domains (the same as FC09).

The proportion of the size of the largest connected component in the object link graph is a little increasing during the last year (from 87.91% to 88.13%), which indicates that the trend of connectivity is to be better slightly.

Effective Diameter. The approximate effective diameter of the OLG from FC08 is 12.28, which is larger than the one from FC09 (11.53). As this index is a highly-accurate approximation, it is likely that the diameter of the OLG shrinks in the past year.

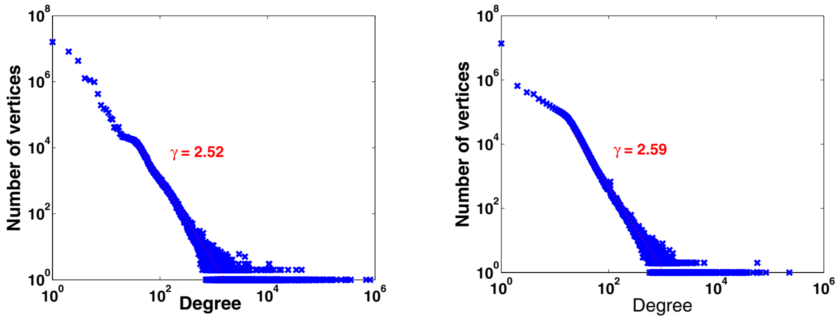
6 Domain-Specific Object Link Structures in the Semantic Web

To verify whether the scale-free nature and small effective diameter also hold for some domain-specific OLGs, and to find the reason why OLG has a sparse structure compared to the traditional Web, we analyze two OLGs constructed from the two largest data sources in FC09, namely Bio2RDF(FC09) and DBpedia(FC09). That is, we construct each OLG from a set of dereference documents from some certain domain (`bio2rdf.org` or `dbpedia.org`). Here, the two OLGs are called the Bio2RDF OLG and the DBpedia OLG respectively.

6.1 Degree

The average degrees of the Bio2RDF OLG and the DBpedia OLG are 3.56 and 3.49 respectively, which indicate a low density in object links in both Bio2RDF(FC09) and DBpedia(FC09). As the homepage of DBpedia[6] points out, DBpedia contains "807,000 links to images and 3,840,000 links to external web pages; 4,878,100 external links into other RDF datasets". In fact, DBpedia functions as a linking-hub for interconnecting various data sources to form the main part of the current Data Web. So its divergent structure results in a low density. As Bio2RDF [4] tries to make documents from public bioinformatics databases, such as Kegg, PDB, MGI, HGNC and several of NCBIs databases, available in

[6] `http://dbpedia.org/`



(a) Bio2rdf.

(b) DBpedia.

Fig. 5. Degree distributions of domain-specific OLGs

RDF format, it contains many small data sources. These small data sources are figured out in the middle-lower part of the Linking Open Data graph.[7] So it is likely the reason that the Bio2RDF OLG has a low density.

Because Bio2rdf(FC09) and DBpedia(FC09) contain 61.13% RDF documents of the total in FC09, it is reasonable to believe that the low density of both Bio2rdf(FC09) and DBpedia(FC09) has great influence on the low density of the object links in FC09.

The number of isolated vertices in the Bio2rdf OLG and the DBpedia OLG are 2,073,663 (5.75%) and 7,193 (0.04%) respectively. Degree distributions of these two OLGs are depicted in Fig. 5. Both two distributions approximately follow power law with exponent 2.52 and 2.59 respectively.

6.2 Connectivity

Connected Component. The largest connected component of the Bio2rdf OLG takes 32,354,360 (89.78%) objects, and the one of DBpedia is 16,499,512 (94.75%). We notice that largest connected components of these two OLGs are both contained in the largest connected component of OLG from FC09, which indicates most objects in these two datasets are interlinked. Besides, both OLGs have a better connectivity than the OLG from FC09 (88.13%).

Effective Diameter. The approximate effective diameters of the Bio2rdf OLG and the DBpedia OLG are 7.38 and 7.81 respectively. Both of them are small as compared to the scale of the graph.

7 Related Work

Graph analysis has been extensively performed on page link graph to the hypertext Web. Albert, et al. [3] analyzed the distributions of incoming and outgoing

[7] `http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-07-14.html`

links between HTML documents in the World Wide Web, and observed power law tails. Adamic [1] showed that the largest strongly connected component of the graph of sites in the Web is a small world. He also counted how many links the sites received from other sites, and found that the distribution of links also follows a power law [2]. Broder, et al. [6] confirmed power law distributions of in- and out-degree. They studied the directed and undirected connected components of the Web, and showed that power laws also arise in the distribution of sizes of those connected components. They revealed that the true structure of the web graph must be somewhat subtler than a "small world" phenomenon in which a browser can pass from any web page to any other with a few clicks. Further, they figured out a bow-tie structure as the macroscopic structure of the Web. Even recently, researcher were still studying various datasets to investigate topological properties of the Web graph, such as bipartite cores, PageRank values, and some correlations [12].

Graph analysis techniques have also been applied to single ontologies or a set of ontologies. Hoser, et al. [16] illustrated the benefits of applying social network analysis to ontologies by measuring SWRC and SUMO ontologies. They interpreted an ontology as a graph: classes and properties became vertices in the graph; an arc was added from a class to its superclass, or from a property to its domain, range, and superproperty. They discussed how different notions of centrality (degree, betweenness, eigenvector, etc.) describe the core content and structure of an ontology, and compared ontologies in size, scope, etc. Zhang [24] studied NCI-Ontology, Full-Galen, and other five ontologies, and discovered that the degree distributions of these entity networks fit power laws well. Theoharis, et al. [21] analyzed graph features of 250 ontologies. For each ontology, they constructed a property graph and a class subsumption graph. The property graph is a directed graph whose vertices correspond to classes and literal types, and whose arcs point from the domain of a property to its range. They found that the majority of ontologies with a significant number of properties approximate a power law for total-degree distribution, and each ontology has a few focal classes that have numerous properties and subclasses. Gil, et al. [13] combined ontologies from the DAML Ontology Library into a single RDF graph, which included 56,592 vertices and 131,130 arcs. They observed that the graph is a small world with an average path length 4.37, and the cumulative degree distribution follows a power law with exponent $\gamma = 1.485$. Recently, Cheng and Qu [9] studied the graph structures of dependence between concepts and between vocabularies. The graphs analyzed in the experiments are constructed from a large dataset that contains more than 1 million terms in more than 3 thousand vocabularies. The results characterize the current status of schemas in the Semantic Web in many aspects, including degree distributions, reachability, and connectivity.

The Semantic Web has also been analyzed from other aspects. Hausenblas, et al. [14] attempted to answering the question: *What is the size of the Semantic Web?* through analyzing the Linking Open Data dataset. Wang, et al. [23] surveyed nearly 1,300 ontologies and analyzed their expressiveness, the use of OWL constructs, the shape of class hierarchy, etc. Tummarello, et al. [22] found that

Table 5. Summary of the characteristics of the OLGs from the four different datasets

dataset	#vertices	#edges	average degree	isolated vertices	γ	largest connected component	effective diameter
FC08	64,974,423	109,373,275	3.37	6.14%	2.74	87.91%	12.28
FC09	110,507,074	190,201,590	3.44	4.29%	2.84	88.13%	11.53
Bio2RDF(FC09)	36,036,245	64,207,033	3.56	5.75%	2.52	89.78%	7.38
DBpedia(FC09)	17,414,639	30,415,886	3.49	0.04%	2.59	94.75%	7.81

the distribution (reuse) of URIs over documents follows a power law. Ding and Finin [10] collected 1,448,504 RDF documents and focused on the distribution of documents over hosts and the sizes of documents. They measured the complexity of terms by counting the number of RDF triples used to define them, and measured the instance space by counting the meta-usages of terms. Power laws were observed in both experiments. Ding, et al. [11] collected over 1.5 million of FOAF documents, and analyzed the empirical usage of namespace and properties in the FOAF community. The authors selected about 7,000 FOAF documents containing 50,559 instances of foaf:Person, and analyzed the social networks induced by those FOAF documents and revealed some interesting patterns. To the best of our knowledge, the macrostructure of the instance level of the Semantic Web has not yet been well studied.

8 Conclusion

In summary, the main contributions of this paper are as follows.

1. A notion of object link graph is proposed to model the Semantic Web structure at the instance level. Based on this notion, we construct an object link graph from a large dataset, namely FC09. We show that the object link graph has the scale-free nature and the effective diameter of the graph is 11.53, which is a small one compared to the scale of the graph.
2. We repeat the complex network analysis on another dataset, namely FC08. The results confirm that the object link graph is scale-free and its effective diameter is a small one compared to the graph's scale. Comparing the object link structure of FC09 with the one of FC08, we observe that the object link graph is not becoming sparser and its diameter is likely to shrink in the past year, though the amounts of documents and objects both doubled, which indicates a good evolution of the Data Web.
3. We repeat the complex network analysis on the two largest domain-specific subsets of FC09, namely Bio2RDF(FC09) and DBpedia(FC09). The results show that both Bio2RDF(FC09) and DBpedia(FC09) have low density in object links, which has great influence on the density of OLG from FC09.

The resulting characteristics of these graphs are summarized in Table 5. The dataset, analyzed graphs, and statistical results are available online.[8] These

[8] `http://ws.nju.edu.cn/olg/`

experimental results presented in this paper can indicate the current state of the object link structure in the Semantic Web.

From these results, we obtain some observations. Firstly, the object link graph inherits some characteristics of the hypertext link structure, such as the scale-free nature. Secondly, a low average degree makes the object link graph different from the page link graph. In fact, a low average degree indicates the lack of links between objects in the Semantic Web, making the object link graph be more sparse, compared with the page link graph. We believe that publishing more object links online will make the Semantic Web better. Besides, more effort is needed to investigate the URI alias phenomena [17] and study the impact of the URI alias phenomenon on the object link structure.

Hopefully, the macroscopic properties of object link structure provided by this paper can help people better understand the current Web of data. In future work, more experiments are deserved to detail the big picture of the object link graph in the Semantic Web, and the dynamic model of the object link graph in the Semantic Web needs to be investigated. Besides, the interaction behavior between instance level and schema level in the Semantic Web is an interesting topic to be studied.

Acknowledgments

The work is supported in part by the NSFC under Grant 60973024, and in part by the NSFC under Grant 60773106. We are grateful to anonymous reviewers for their precious comments.

References

1. Adamic, L.A.: The Small World Web. Research and Advanced Technology for Digital Libraries 1696, 443–452 (1999)
2. Adamic, L.A., Huberman, B.A.: Power-Law Distribution of the World Wide Web. Science 287(5461), 2115a (2000)
3. Albert, R., Jeong, H., Barabasi, A.L.: The Diameter of the World Wide Web. Nature 401, 130–131 (1999)
4. Belleau, F., Nolin, M., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. Journal of Biomedical Informatics 41(5), 706–716 (2008)
5. Berners-Lee, T.: Linked Data - Design Issues (2006), `http://www.w3.org/DesignIssues/LinkedData.html`
6. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer Networks 33(1-6), 309–320 (2000)
7. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law Distributions in Empirical Data. SIAM Review 51, 661–703 (2009)
8. Cheng, G., Ge, W., Qu, Y.: Falcons: Searching and Browsing Entities on the Semantic Web. In: Proceedings of the 17th International Conference on World Wide Web, pp. 1101–1102 (2008)

9. Cheng, G., Qu, Y.: Term dependence on the semantic Web. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 665–680. Springer, Heidelberg (2008)
10. Ding, L., Finin, T.: Characterizing the Semantic Web on the Web. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 242–257. Springer, Heidelberg (2006)
11. Ding, L., Zhou, L., Finin, T., Joshi, A.: How the Semantic Web is Being Used: An Analysis of FOAF Documents. In: Proc. of 38th Annual Hawaii International Conference on System Sciences, p. 113c (2005)
12. Donato, D., Laura, L., Leonardi, S., Millozzi, S.: The Web as a graph: How far we are. ACM Transactions on Internet Technology (TOIT) 7(1) (2007)
13. Gil, R., García, R., Delgado, J.: Measuring the Semantic Web. AIS SIGSEMIS Bulletin, 69–72 (2004)
14. Hausenblas, M., Halb, W., Raimond, Y., Heath, T.: What is the Size of the Semantic Web? In: Proc. of I-Semnatics (2008)
15. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D.: Web science: an interdisciplinary approach to understanding the web. Communications of the ACM 51(7), 60–69 (2008)
16. Hoser, B., Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Semantic Network Analysis of Ontologies. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 514–529. Springer, Heidelberg (2006)
17. Jacobs, I., Walsh, N. (eds.): Architecture of the world wide web, volume one. W3C Recommendation December 15 (2004), `http://www.w3.org/TR/webarch/`
18. Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Extracting Large-scale Knowledge Bases from the Web. In: International Conference on Very Large Data Bases, pp. 639–650 (1999)
19. Leskovec, J., Kleinbery, J., Faloutsos, C.: Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In: Proc. of International Conference on Knowledge Discovery and Data Mining, pp. 177–187 (2005)
20. Palmer, C.R., Gibbons, P.B., Faloutsos, C.: ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs. In: Proc. of International Conference on Knowledge Discovery and Data Mining, pp. 81–90 (2002)
21. Theoharis, Y., Tzitzikas, Y., Kotzinos, D., Christophides, V.: On Graph Features of Semantic Web Schemas. IEEE Transactions on Knowledge and Data Engineering 20(5), 692–702 (2008)
22. Tummarello, G., Delbru, R., Oren, E.: Sindice.com:Weaving the Open Linked Data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 552–565. Springer, Heidelberg (2007)
23. Wang, T.D., Parsia, B., Hendler, J.: A Survey of the Web Ontology Landscape. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 682–694. Springer, Heidelberg (2006)
24. Zhang, H.: The Scale-Free Nature of Semantic Web Ontology. In: Proc. of the 17th International Conference on World Wide Web, pp. 1047–1048 (2008) (poster)