How Matchable Are Four Thousand Ontologies on the Semantic Web

Wei Hu, Jianfeng Chen, Hang Zhang, and Yuzhong Qu

State Key Laboratory for Novel Software Technology, Nanjing University, China
{whu,yzqu}@nju.edu.cn, jf_chen@ymail.com,
hzhang.nju@gmail.com

Abstract. A growing number of ontologies have been published on the Semantic Web by various parties, to be shared for describing things. Because of the decentralized nature of the Web, there often exist different but similar ontologies from overlapped domains, or even within the same domain. In this paper, we collect more than four thousand ontologies and perform a large-scale pairwise matching based on an ontology matching tool. We create about three million mappings between the terms (classes and properties) in these ontologies, and construct a complex term mapping graph with terms as nodes and mappings as edges. We analyze the macroscopic properties of the term mapping graph as well as the derived ontology mapping graph, which characterize the global ontology matchability in several aspects, including the degree distribution, connectivity and reachability. We further establish a pay-level-domain mapping graph to understand the common interests between different ontology publishers. Additionally, we publish the generated mappings online based on the R2R mapping framework. These mappings and our observations are believed to be useful for the Linked Data community in ontology creation, integration and maintenance.

1 Introduction

The *Semantic Web* is an ongoing effort by the W3C Semantic Web Activity for realizing data integration and sharing across different applications and parties. As of today, a growing number of popular *ontologies* have emerged to describe things for specific domains, e.g., the Friend of a Friend (FOAF). These ontologies recommend common classes and properties (uniformly called *terms* in this paper) that are widely and consistently used in data sources.

Because of the decentralized nature of the Web, there usually exist multiple ontologies from overlapped application domains or even within the same domain. In order to establish interoperability between (Semantic) Web applications that use different but related ontologies, *ontology matching* (OM) has been proposed as an effective way for handling the semantic heterogeneity problem. It is useful for many tasks, such as data integration and distributed query processing.

To date, a large amount of (semi-)automatic OM approaches have been proposed in literature [10], which exploit a wide range of characteristics in ontologies,

G. Antoniou et al. (Eds.): ESWC 2011, Part I, LNCS 6643, pp. 290–304, 2011.

such as linguistic descriptions, structures, data instances, and even background knowledge from thesaurus or third parties' ontologies. But, the global analysis on ontology matchability is still missing, that is, *how matchable are the ontologies on the Semantic Web so far?* In this paper, we dedicate to answering this question. We believe that the study on the morphology of ontology overlaps is important for the Semantic Web, and our observations would help ontology developers and users in the process of ontology creation, integration and maintenance.

Complex network analysis has been widely performed on the page link graph to investigate some macroscopic properties of the Hypertext Web [1,2,5,9]. Recently, such analysis techniques have been applied to the Semantic Web as well, from small sets of ontologies [12,14,18] to the large Linked Data cloud [8,11,17]. However, to the best of our knowledge, because of the high computational cost, the macroscopic matchability among ontologies on the whole Semantic Web has not been well studied yet.

In this paper, we collect more than four thousand Web ontologies by a Semantic Web search engine named Falcons [6], and employ six computers running nearly a year to perform a large-scale pairwise matching by an ontology matching tool named Falcon-AO [16]. We create about 3.1 million mappings between two million terms from the ontologies, and build a complex *term mapping graph*, where nodes are derived from terms and edges are from mappings.

Then, we analyze the macroscopic properties of the term mapping graph in many aspects, including the degree distribution, average distance and clustering coefficient. In addition, we derive an *ontology mapping graph*, in which directed edges are derived from the term mappings with respect to the size of ontologies, for analyzing how big the overlaps of these ontologies are. According to our experiment, we observe that, both the term mapping graph and ontology mapping graph exhibit the scale-free nature with a few "hubs", and the terms (ontologies) from a large part of the graphs form a small world.

Furthermore, we categorize the ontologies in terms of the pay-level-domains of their namespaces (a *pay-level-domain mapping graph*), and observe the common interests among various ontology publishers. We see that `dbpedia.org` and `umbc.edu` are two generic ontology publishers and their ontologies cover a broad range of real-world domains. In addition, our created mappings are all published online[1] based on the R2R mapping specification [4], which would facilitate the Linked Data community to create, integrate and maintain ontologies. Moreover, we are trying to apply these mappings to enhance our Semantic Web browser[2] by recommending matchable terms to general users.

The rest of this paper is organized as follows. Sect. 2 discusses related work. The dataset, metrics and tools used in the experiment are introduced in Sect. 3. In Sect. 4 and Sect. 5, we analyze the macroscopic properties of the term mapping graph and the derived ontology mapping graph, respectively. We further at a higher level investigate the pay-level-domain mapping graph in Sect. 6. Finally, Sect. 7 summarizes our findings in this paper and points out future work.

[1] `http://ws.nju.edu.cn/mappings/`

[2] `http://ws.nju.edu.cn/explorer/`

2 Related Work

Graph analysis has been extensively studied on page link graphs for the Hypertext Web. Albert, et al. [2] analyzed the distributions of incoming and outgoing links between HTML documents on the Web, and observed the power law tails. Adamic and Huberman [1] observed the small world phenomenon in the largest strongly connected component of the website graph. Broder, et al. [5] confirmed the power law distributions of in-degrees and out-degrees, and discovered that a power law also appears in the distribution of the sizes of connected components. They figured out a "bow-tie" structure as the macroscopic structure of the Web. Even recently, researchers were still studying various datasets to investigate the topological properties of the Web [9].

Graph analysis techniques have been conducted to a single ontology or a set of ontologies. Hoser, et al. [15] illustrated some benefits of applying social network analysis to the SWRC and SUMO ontologies, and discussed how different notions of centrality (e.g., degree, betweenness, eigenvector) describe the core content and structure of an ontology. Theoharis, et al. [18] observed the graph features of 250 ontologies, and claimed that a majority of ontologies with a significant number of properties approximate power laws for the total-degrees, and each ontology has a few focal classes with numerous properties and subclasses. At a larger scale, Gil, et al. [13] combined the ontologies from the DAML ontology library into a single RDF graph, which includes 56,592 nodes and 131,130 edges. They found that this graph is a small world and the cumulative degree distribution follows a power law. Tummarello, et al. [21] observed that the distribution (reuse) of URIs over documents follows a power law. Ding, et al. [8] gave a quantitative analysis of `owl:sameAs` deployment status and used these statistics to focus discussion around its usage in Linked Data.

To the best of our knowledge, there are two works that address the analysis of ontology matchability. Ghazvinian, et al. [12] investigated the morphology of ontology mappings among 207 biomedical ontologies, where the mappings were extracted by similar names. Nikolov and Motta [17] created term mappings from declared coreference association (e.g., `owl:sameAs`) and co-typing, and analyzed a snapshot of the Billion Triple Challenge 2009 containing several hundreds of ontologies. Their mappings hold not only the equivalence relations but also the subsumption, e.g., `movie:actor` co-types with `foaf:Person`. In this paper, we analyze over four thousand ontologies, which is much larger than the sizes of the two previous works. Additionally, we use a general ontology matching approach which does not tailor itself to some specific domains. Our experimental results also show some different observations as compared with [12,17].

3 Experiment Setting

The goal of our work is to investigate the macroscopic matchability of ontologies in a dataset that contains a significant number of mappings. We therefore introduce the notion of ontologies, metrics in the experiment and the ontology matching tool Falcon-AO.

3.1 Statistical Data of Ontologies

An ontology $\mathcal{O}$ is viewed as a triple $\langle id, \mathcal{V}, \mathcal{G} \rangle$, where *id* is a unique identifier; $\mathcal{V}$ is a vocabulary that consists of a non-empty set of terms (classes and properties) holding a common URI namespace [3]; and $\mathcal{G}$ is an RDF graph that describes the terms in $\mathcal{V}$.

We recognize all the vocabularies and their involved terms from Falcons, and dereference their URIs to obtain the dereferenced documents. The identifier of an ontology is the URI namespace of its vocabulary, and the RDF triples in all dereferenced documents are merged as the RDF graph of that ontology, because the dereferenced documents for a vocabulary and its involved terms could be different. Terms having the same namespace as $\mathcal{O}$ are called *local* terms, others are referred to as *external* ones.

Based upon a snapshot of the Semantic Web data collected by Falcons until September 2009, we collect 4,433 Web ontologies, in which most are written in RDF(S) and OWL, while merely a small amount are in DAML+OIL. It is worth noting that, if we define ontologies with respect to separately stored dereferenced documents rather than merging them together, the number of ontologies would be about 25 thousands.

The ontologies contain 2,033,935 local terms in total that cover a lot of real-world domains, e.g., social community, academic publication, music, movie and geography. More specifically, the terms can be classified into 1,895,030 classes and 138,905 properties. A few ontologies have extremely large number of terms, such as YAGO, Cyc, ETHAN, DBpedia and biomedical ontologies FMA, Gene and MeSH. The distribution of the number of terms per ontology is illustrated in Fig. 1, which indicates that the distribution approximates a power law with the exponent $\beta = 1.34$. This power law distribution is in accordance with the observation in [21].

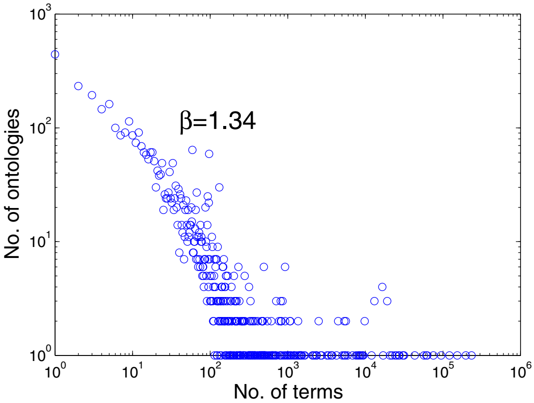


Fig. 1. Power law distribution of the number of ontologies versus the number of terms per ontology

3.2 Experimental Metrics

A *graph* G consists of a finite, non-empty set of *nodes* N and a set of *edges* E. An edge in E is an ordered (for directed graphs) or an unordered (for undirected graphs) pair (u, v), which denotes a connection between $u \in N$ and $v \in N$.

A *weakly/strongly connected component* of G is a subgraph in which any two nodes can be reachable to each other through undirected/directed paths, and to which no more nodes or edges can be added while still preserving its reachability. The number of nodes in a connected component is called its *size*.

The *average distance* for a connected graph is measured as the average shortest path lengths between all the nodes in it. The local clustering coefficient [22] for a node in a connected graph quantifies how close its neighbors are to be a clique (complete graph), and the *clustering coefficient* for the graph is the average of the local clustering coefficients of all nodes. A graph exhibits the *small world* phenomenon, if its clustering coefficient is significantly higher than that of a random graph on the same node set, and if the graph has a short average distance.

A random variable x is distributed according to a *power law* when its probability density function $p(x)$ is in the form of $p(x) = \alpha x^{-\beta}$, where α, β are positive constants, and β is called the *power law exponent*. Power law functions are scale-free, in the sense that if x is re-scaled by multiplying it by a constant, $p(x)$ would still be proportional to $x^{-\beta}$ [18]. According to [7], β can be estimated based on a maximum-likelihood method as follows:

$$\beta \approx 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1} . \tag{1}$$

3.3 Ontology Matching and Falcon-AO

Ontology matching (also called mapping or aligning) aims at creating mappings (also known as alignments, correspondences or matches) between semantically matchable terms from different ontologies [10]. In this paper, we define that a *term mapping* is constituted by two terms that hold an equivalence relation, and the matchability between them is in $(0, 1]$ range.

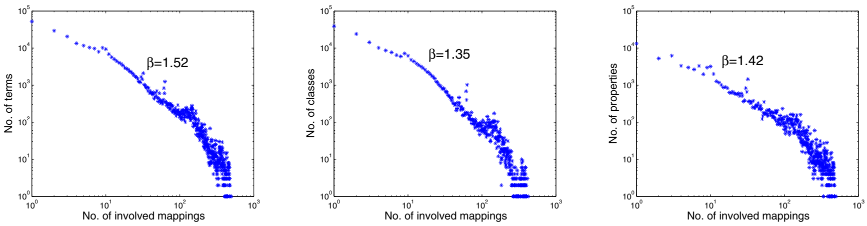
Falcon-AO [16] is a generic, automatic OM tool, which accepts as input two ontologies to be matched, and supplies a library of the edit-distance based and TF-IDF based matchers, the similarity propagating matcher and the partition-based block matcher for large ontologies. Falcon-AO was one of the best tools in all kinds of tests in the OAEI campaign from 2005 to 2007, including the Benchmark, Conference, Directory, Anatomy, Food and Library tracks. Besides good performance, the reasons for selecting Falcon-AO in our experiment include: (i) Falcon-AO is scalable, which is feasible to match very large ontologies; (ii) it is open source. We can easily fix exceptions/bugs during matching; and (iii) it can be run in a batch mode.

4 Term Mapping Graph Analysis

We employ six personal computers and spend nearly one year to pairwise match those 4,433 ontologies, which is a time-consuming process. We create approximate 6 million term mappings with Falcon-AO. In order to make the following analysis more convincible, we filter the mappings with matchability less than 0.7. According to our past experience in OAEI, the threshold 0.7 indicates that the remained mappings can be of high-precision. We also randomly choose a set of 5,000 mappings and perform manual judgement on them. The average precision is about 0.965. After filtering, we retain 3,099,393 mappings between terms. Some statistical data are as follows.

1. Only 280,733 local terms (195,669 classes and 85,064 properties) are involved in these mappings, which are a small part with respect to the total number of terms (2,033,935) in all the ontologies. It indicates that 86.2% terms are unique on the Semantic Web.
2. The number of mappings between classes is 1,553,740 and the number between properties is 1,545,653. In average, a class participates in about 7.9 mappings, while a property is in 18.2 mappings, indicating that properties are more matchable than classes.
3. 45.6% (1,414,406) mappings involve terms with different local names. The local name of a term is a string after the last hash "#" or slash "/" of its URI. There exist one-to-many mappings even within a pair of ontologies, namely, one term in one ontology might be matchable with more than one terms in the other ontology. We totally find 69,457 terms that participate in one-to-many mappings.

Based upon the 3.1 million mappings, we can establish a term mapping graph, where edges are derived from the mappings and nodes are from the terms involved in these mappings. The term mapping graph is undirected, because the mappings that we generate are symmetric. In addition, classes are only matched with classes while properties are matched with properties, so we separately construct a class mapping graph and a property mapping graph.

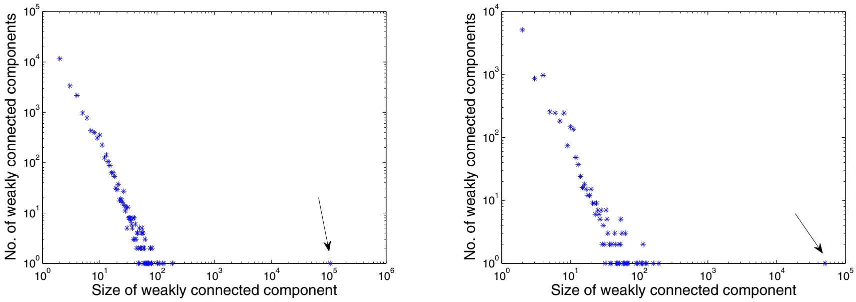


(a) Terms (b) Classes (c) Properties

Fig. 2. Power law distributions of the number of terms, classes and properties versus the number of involved mappings per term, class and property

Fig. 2(a) shows that the distribution of the number of terms versus the number of involved mappings per term approximates a power law with the exponent $\beta = 1.52$. The distribution does not depict a long tail, which indicates the moderate number of mappings for each term, in other words, no term matches an extremely large amount of other terms. Analogously, the distributions for classes (see Fig. 2(b)) and properties (see Fig. 2(c)) also approximate power laws without long tails. Due to the decentralized nature of the Semantic Web, everyone can publish their own ontologies, which results in many heterogenous definitions of common terms that constitute mappings. As time goes on, some collections of well-defined terms outperform other ones and are universally accepted. So, nonexistence of the long tail reveals the evolution process of ontologies on the Semantic Web.

Fig. 3(a) and 3(b) illustrate the distributions of the number of weakly connected components versus the size of weakly connected component for classes and properties, respectively. Most of these weakly connected components, excluding the largest ones, have sizes less than 200. In addition, the clustering coefficient of the largest weakly connected component for classes is 0.601 and for properties is 0.719, while the average distance for classes is 19.28 and for properties is 8.81, which demonstrate that the property mapping graph forms a small world, however the class mapping graph does not. The average distance for classes is larger than that for properties, which is in accordance with the result that the mappings between classes are sparser than those between properties.



(a) Classes

(b) Properties

Fig. 3. Distribution of the number of weakly connected components versus the size of weakly connected component

Most weakly connected components have moderate sizes, but the two largest weakly connected components for classes and properties are so large that we need to conduct a deeper investigation. We find that, in contrast to the small weakly connected components, the matchable terms in the largest ones are not so equivalent to each other. This phenomenon can be interpreted as a result of mapping composition [19], caused by ontology or term characteristics deviating in meaning. Then, after merging mappings into a graph, several clusters different in semantics are bridged by some unreasonable mappings, resulting in huge weakly connected components. Therefore, it gives a lesson that we cannot heavily believe in the mapping chains between terms, especially when the transitive chains are long, due to wrong term mapping composition.

Furthermore, we also investigate the most popular local names for classes or properties. We extract the local name of each class or property in our mappings, and count the times of each local name appears (by ignoring string cases). The top-5 local names for classes and properties are listed in Table 1(a) and 1(b), respectively. We see that, although some terms like `foaf:Person` or `dc:title` have been widely accepted, they are still duplicately defined with different URIs in many ontologies. For instance, there are `dbpedia:Person`, `umbel:Person` and many others. These legacy duplications may cause some difficulties in data sharing and reuse, since they weaken the network effect of the Semantic Web.

Table 1. Top-5 popular local names for classes and properties

(a) Classes

	Local name	Times
1	Person	372
2	Organization	254
3	Book	213
4	Article	204
5	Address	179

(b) Properties

	Local name	Times
1	name	468
2	title	278
3	type	237
4	location	237
5	date	205

5 Ontology Mapping Graph Analysis

In this section, we firstly describe the notion of directed edges between ontologies, and then analyze the macroscopic properties of the derived ontology mapping graph.

5.1 Construction of Edges between Ontologies

Each node in an ontology mapping graph is derived from an ontology, while each directed edge is from a set of term mappings between two ontologies. Because the transitivity of matchability in a term mapping graph may cause problems, constructing an edge between two ontologies just depends on those mappings between the terms located in the two ontologies. In other words, an edge exists between two ontologies iff there are explicit term mappings between them.

Although the mappings between terms are undirected, edges between ontologies need a more proper definition, since an ontology contains a collection of terms. Assuming that we have created several mappings between two ontologies, e.g., Food and Pizza. The number of terms involved in the mappings is nearly the same, but considering people's intuitions the matchability is not symmetric especially when noticing the disparity in the sizes of these two ontologies. We may say that Pizza is more matchable to Food whereas the matchability decreases in the other direction. This is supported by the Tversky contrast model

[20], which proposed to compute asymmetric matchability by taking into account both common and different "features" of the things being compared.

Based on the intuition mentioned above and also inspired by [12], we propose a percent-normalized directed edge between two ontologies, which considers not only the term mappings between ontologies, but also the sizes of ontologies for direction and normalization.

Let $\mathcal{O}$ be the source ontology and $\mathcal{O}'$ be the target ontology. $\mathcal{M}^{\gamma}_{\mathcal{O},\mathcal{O}'}$ denotes a set of mappings between the terms in $\mathcal{O},\mathcal{O}'$ holding their matchability $\geq \gamma$, where $\gamma \in [0.7, 1)$. $\mathcal{T}(\mathcal{O})$ denotes all the local terms in $\mathcal{O}$. $\mathcal{I}(\mathcal{O}, \mathcal{M}^{\gamma}_{\mathcal{O},\mathcal{O}'}) = \{t \in \mathcal{T}(\mathcal{O}) \mid \exists \langle t, t' \rangle \in \mathcal{M}^{\gamma}_{\mathcal{O},\mathcal{O}'}\}$, representing the local terms in $\mathcal{O}$ that are *involved* in $\mathcal{M}^{\gamma}_{\mathcal{O},\mathcal{O}'}$. A directed edge $e^{\gamma,q}_{d}$ from $\mathcal{O}$ to $\mathcal{O}'$ exists iff $match^{\gamma}_{d}(\mathcal{O},\mathcal{O}') > q$, where $q \in [0, 1)$ and $match^{\gamma}_{d}()$ is defined for measuring how big the overlaps of terms between two ontologies:

$$match^{\gamma}_{d}(\mathcal{O},\mathcal{O}') \triangleq \frac{|\mathcal{I}(\mathcal{O}, \mathcal{M}^{\gamma}_{\mathcal{O},\mathcal{O}'})|}{|\mathcal{T}(\mathcal{O})|}. \tag{2}$$

The directed percent-normalized edges reveal how significantly a set of term mappings affect the matchability between ontologies. By changing q, we analyze the characteristics of ontology mapping graph.

5.2 Results

Fig. 4 depicts the variation of the number of edges for different values of q. More specifically, when $q = 0$, the created ontology mapping graph has 1,618,330 directed edges. But the number of edges sharply falls with the increase of q. As compared with the number of ontology pairs, i.e., 9,823,528 ($4433^2/2$), the quantity of edges is quite rare. However, if we realize that the ontologies are collected from the Web and diverse in domains, this result makes sense.

Fig. 5 illustrates the variation of several graph features with the values of q changing from 0.0 to 0.9 for the ontology mapping graph, including the number of connected (not isolated) ontologies, the number of weakly connected components,

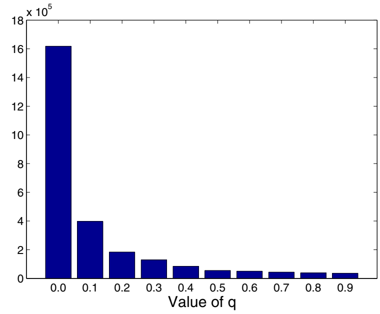


Fig. 4. Variation of the number of directed edges with different values of q

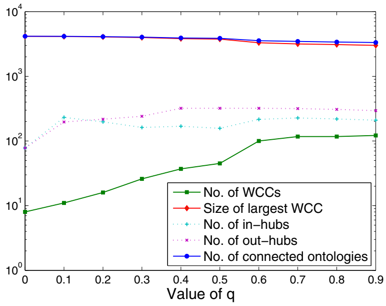
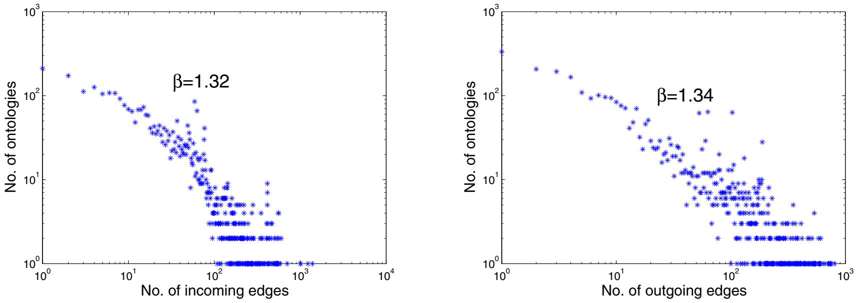


Fig. 5. Variation of graph features with different values of q

the size of the largest weakly connected component and the number of hubs. Here we focus on two kinds of hubs: (i) an in-hub has more than twice the average number of incoming edges of nodes, and (ii) an out-hub has more than twice the average number of outgoing edges of nodes.

With the increase of q, there are more and more isolated ontologies and the size of the largest weakly connected component is shrinking. But the changes are not very sharp and the weakly connected component is almost as large as the whole ontology mapping graph, which indicates a very strong matchability between ontologies. Besides, the proportion of in/out-hubs almost keeps at the level of 10%, which reveals that a small portion of ontologies take part in the connectivity of ontology mapping graph. For $q = 0.1$, the clustering coefficient of the largest weakly connected component is 0.403, while the average distance between the ontologies in it is 2.3, which forms a small world. But, this distance is larger than the one (1.1) in [12], indicating that the average distance between the ontologies in our weakly connected component is longer than that of the particular biomedical domain.

Under a higher threshold value $q = 0.95$ for determining two ontology are matchable or not, we also observe two interesting phenomena. One is the versional evolution which produces a series of versions with slight changes for the same ontology, while the other is the duplicate deployment of the same ontology under different namespaces. Moreover, these two cases often mix with each other. For example, we find two different versions of the Pizza ontology whose version ID changes from 1.1 to 1.4. This ontology is also copied with dozens of different namespaces.



(a) Incoming (b) Outgoing

Fig. 6. Power law distribution of the number of ontologies versus the number of incoming/outgoing edges per ontology under $q = 0.2$

Due to space limitation, we merely show here the distributions of the number of incoming and outgoing edges per ontology under $q = 0.2$ in Fig. 6(a) and 6(b), respectively. Both the distributions follow power laws. It is interesting to note that, when q is set to other values (e.g., 0.1, 0.3 or 0.4), the distributions still approximate power laws. Such scale-free nature tells that a few prominent ontologies dominate the connectively of the ontology mapping graph.

The top-5 ontologies with most incoming and outgoing edges for $q = 0.2$ are listed in Table 2 and 3, respectively. Referring to the number of terms contained by the ontologies, in-hubs are usually those ontologies large in size while out-hubs are usually those ontologies small in size. We also observe that in-hubs usually represent common knowledge bases such as DBpedia or prominent ontologies in the mature domains on the Semantic Web, e.g., DCD from the field of biomedicine. [12] indicates that hubs with many outgoing edges show shared domains, in particular at high threshold values for q. However, this method is ineffective for identifying shared domains on the whole Web, because ontologies from the Semantic Web have a great diversity in their sizes and other fields are not as mature as the biomedicine field.

Table 2. Top-5 ontologies with most incoming edges under $q = 0.2$

	URI	#Edges	#Terms
1	`http://dbpedia.org/property/`	1389	24215
2	`http://www.cs.umbc.edu/~aks1/ontosem.owl#`	1228	8501
3	`http://athena.ics.forth.gr:9090/RDF/.../DCD100.rdf#`	1008	5354
4	`http://dbpedia.org/ontology/`	706	889
5	`http://counterterror.mindswap.org/2005/terrorism.owl#`	602	501

Table 3. Top-5 ontologies with most outgoing edges under $q = 0.2$

	URI	#Edges	#Terms
1	`http://vistology.com/ont/bug/error/owl/person.owl#`	809	5
2	`http://www.vistology.com/ont/tests/student4.owl`	757	5
3	`http://tbc.sk/RDF/entity.rdf#`	746	5
4	`http://vistology.com/ont/tests/owlError1.owl#`	737	4
5	`http://vistology.com/.../similarUnused/.../person.owl#`	724	5

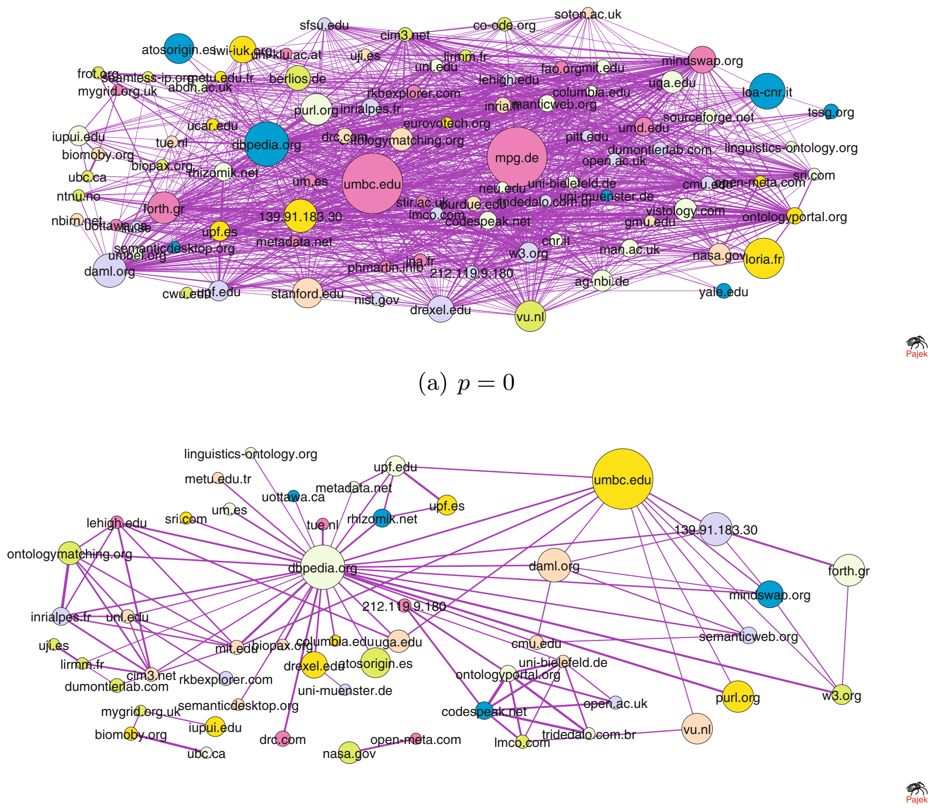
6 Pay-Level-Domain Mapping Graph Analysis

During matching ontologies on the Semantic Web, some common interests between different ontology publishers can be distilled. To reveal relations between these publishers, we introduce the pay-level-domain mapping graph to categorize ontologies into different pay-level-domains and identify their relations.

A pay-level-domain mapping graph is defined as an undirected graph, where each node denotes a pay-level-domain that is constituted by the ontologies belonging to it, while each edge denotes a relation between two domains. Let $\mathcal{D}, \mathcal{D}'$ be two pay-level-domains. $\mathcal{M}^{\eta}_{\mathcal{D},\mathcal{D}'}$ denotes a set of mappings among the ontologies in $\mathcal{D}, \mathcal{D}'$ with their matchability $\geq \eta$, where $\eta \in [0, 1)$. $\mathcal{O}(\mathcal{D})$ gives all the ontologies in $\mathcal{D}$. $\mathcal{J}(\mathcal{D}, \mathcal{M}^{\eta}_{\mathcal{D},\mathcal{D}'}) = \{o \in \mathcal{O}(\mathcal{D}) \mid \exists \langle o, o' \rangle \in \mathcal{M}^{\eta}_{\mathcal{D},\mathcal{D}'}\}$, denoting the ontologies in $\mathcal{D}$ that are involved in $\mathcal{M}^{\eta}_{\mathcal{D},\mathcal{D}'}$. An undirected edge $e^{\eta,p}_{u}$ between

$\mathcal{D}, \mathcal{D}'$ exists iff $match_u^\eta(\mathcal{D}, \mathcal{D}') > p$, where $p \in [0, 1)$ and $match_u^\eta()$ is defined to indicate how big the overlaps of ontologies between two publishers:

$$match_u^\eta(\mathcal{D}, \mathcal{D}') \triangleq \min\left(\frac{|\mathcal{J}(\mathcal{D}, \mathcal{M}_{\mathcal{D},\mathcal{D}'}^\eta)|}{|\mathcal{O}(\mathcal{D})|}, \frac{|\mathcal{J}(\mathcal{D}', \mathcal{M}_{\mathcal{D},\mathcal{D}'}^\eta)|}{|\mathcal{O}(\mathcal{D}')|}\right). \tag{3}$$



(a) $p = 0$

(b) $p = 0.3$

Fig. 7. Pay-level-domain mapping graphs under two different values of p (only top-100 pay-level-domains are shown with respect to the number of terms per domain)

We use Nutch[3] to obtain 395 pay-level-domains for the 4,433 ontologies, and select ontology mappings that hold their matchability greater than 0.2 to avoid the influence of "noisy" mappings. Under $p = 0$, the pay-level-domain mapping graph generated from Pajek[4] using the top-100 biggest pay-level-domains with respect to the number of terms in each domain is depicted in Fig. 7(a). There are 90 domains matchable with each other, while the left 10 domains have no

[3] http://nutch.apache.org/

[4] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

connection with others, since the ontology mappings for these domains have low matchability and are filtered before. There is only one big connected component in the figure, and the mappings between these domains are very complex, which means that most ontology publishers connect with each other more or less.

We increase p to 0.3 in order to filter some insignificant edges. Many edges are omitted and only 95 edges left. As a result, 38 domains are removed since all edges linking to them are deleted. A clearer depiction is shown in Fig. 7(b), where `DBpedia.org` is the biggest hub in this pay-level-domain mapping graph and `umbc.edu` ranks the second. Several active organizations on the Semantic Web, e.g., UMBC, W3C and DAML, have already provided a large amount of ontologies, which are matched with other ones. In view of `DBpedia.org`, data producers are welcome to link their data into the Linked Data cloud, thus `DBpedia.org` becomes a central point on the Semantic Web. Besides, two small connected components are separated, where one contains two domains (`nasa.gov` and `open-meta.com`), and the other contains four domains about biomedicine.

The visualization of the pay-level-domain mapping graph provides insights into how publishers are connected. Both Fig. 7(a) and 7(b) show a picture of how publishers share their interests. Note that pay-level-domain mapping graph is based on ontology mapping graph, we can conclude that most publishers are interested in a diversity of topics, which is demonstrated in Fig. 7(a), while deep interests can be seen from a specified pay-level-domain mapping graph (e.g., Fig. 7(b)).

7 Conclusion

In this paper, we collect more than 4 thousand ontologies based on the Falcons search engine, and perform a large-scale pairwise matching with a scalable ontology matching tool Falcon-AO. We generate 3.1 million term mappings, which are used for analyzing the morphology of term mapping graph, ontology mapping graph and pay-level-domain mapping graph. To the best of our knowledge, our work is the first attempt towards analyzing the matchability between such a large number of ontologies on the Semantic Web, where the difficulties lie in the high computational cost and the messiness of real Semantic Web data.

By analysis, we make some observations as follows. Firstly, both the term mapping graph and ontology mapping graph inherit some characteristics of the Hypertext Web and Semantic Web, such as the scale-free nature and the small world. Secondly, a small portion of terms are well matched, while many cannot match any others, which demonstrate the skewed matchability between terms. However, most ontologies are loosely connected to each other. Thirdly, ontology publishers show common interests in ontology development, where `DBpedia.org` and `umbc.edu` are the two most active publishers. Lastly, our experimental results confirm some existing conclusions on a small set of ontologies, but we also find some differences. For example, the average distance between our ontologies is twice larger than the one in the biomedical domain.

From a practical viewpoint, our downloadable term mappings can help both ontology developers and users in the process of ontology creation, integration and maintenance. For example, before creating an ontology, developers could check if similar terms (e.g., `foaf:Person`, `dc:title`) or ontologies have already been defined. Even if ontologies were created, people can still link their terms with popular ones based on our mappings, by using `owl:equivalentClass` and `owl:seeAlso` constructs to gain potential interoperability. We believe that, for some domains, reusing well-known ontologies and terms rather than "reinventing the wheel" would facilitate data integration and sharing for a better Data Web; while for other domains, more efforts are expected to create new ontologies or synthesize existing ones. Another example is when conducting ontology matching or data fusion, users can identify representative hubs (e.g., DBpedia, SUMO and OpenGALEN) as useful background knowledge.

The analytic results reported in this paper is just the first step, and many issues still need to be addressed further. In the near future, we look forward to using other robust and scalable ontology matching tools to repeat some part of the experiment and confirm our observations. Another important problem raised from our study is how to utilize these mappings for potential applications, such as object consolidation and Semantic Web data browsing.

Acknowledgements

This work is supported in part by the NSFC under Grant 61003018, 60973024 and 61021062, and also in part by the National Research Foundation for the Doctoral Program of Higher Education of China under Grant 20100091120041. We are grateful to Gong Cheng and Weiyi Ge for providing ontologies in the Falcons search engine. We also appreciate the helpful comments from three anonymous reviewers.

References

1. Adamic, L., Huberman, B.: Power-Law Distribution of the World Wide Web. Science 287(5461), 2115a (2000)
2. Albert, R., Jeong, H., Barabasi, A.: The Diameter of the World Wide Web. Nature 401, 130–131 (1999)
3. Berrueta, D., Phipps, J.: Best Practice Recipes for Publishing RDF Vocabularies. W3C Working Group Note (2008)
4. Bizer, C., Schultz, A.: The R2R Framework: Publishing and Discovering Mappings on the Web. In: ISWC Workshop on Consuming Linked Data (2010)
5. 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)
6. Cheng, G., Qu, Y.: Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. International Journal on Semantic Web and Information Systems 5(3), 49–70 (2009)
7. Clauset, A., Shalizi, C., Newman, M.: Power-Law Distributions in Empirical Data. SIAM Review 51(4), 661–703 (2009)

8. Ding, L., Shinavier, J., Shangguan, Z., McGuinness, D.: SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl:sameAs in Linked Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 145–160. Springer, Heidelberg (2010)
9. Donato, D., Laura, L., Leonardi, S., Millozzi, S.: The Web as a Graph: How Far We Are. ACM Transactions on Internet Technology 7(1), 1–25 (2007)
10. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
11. Ge, W., Chen, J., Hu, W., Qu, Y.: Object Link Structure in the Semantic Web. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010. LNCS, vol. 6089, pp. 257–271. Springer, Heidelberg (2010)
12. Ghazvinian, A., Noy, N., Jonquet, C., Shah, N., Musen, M.: What Four Million Mappings Can Tell You about Two Hundred Ontologies. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 229–242. Springer, Heidelberg (2009)
13. Gil, R., Garcia, R., Delgado, J.: Measuring the Semantic Web. AIS SIGSEMIS Bulletin, 69–72 (2004)
14. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010)
15. 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)
16. Hu, W., Qu, Y.: Falcon-AO: A Practical Ontology Matching System. Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 237–239 (2008)
17. Nikolov, A., Motta, E.: Capturing Emerging Relations Between Schema Ontologies on the Web of Data. In: ISWC Workshop on Consuming Linked Data (2010)
18. Theoharis, Y., Tzitzikas, Y., Kotzinos, D., Christophides, V.: On Graph Features of Semantic Web Schemas. IEEE Transcations on Knowledge and Data Engineering 20(5), 692–702 (2008)
19. Tordai, A., Ghazvinian, A., van Ossenbruggen, J., Musen, M., Noy, N.: Lost in Translation? Empirical Analysis of Mapping Compositions for Large Ontologies. In: ISWC Workshop on Ontology Matching (2010)
20. Tversky, A.: Features of Similarity. Psychological Review 84(4), 327–352 (1977)
21. 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)
22. Watts, D., Strogatz, S.: Collective Dynamics of 'Small-World Networks'. Nature 393(6684), 440–442 (1998)