Article

Journal of Information Science
1–27

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Characterizing RDF Datasets

Javier D. Fernández

Vienna University of Economics and Business, Vienna, Austria

Miguel A. Martínez-Prieto

Department of Computer Science, University of Valladolid, Valladolid, Spain

Pablo de la Fuente Redondo

Department of Computer Science, University of Valladolid, Valladolid, Spain

Claudio Gutiérrez

Computer Science Department and CIWR, Universidad de Chile, Santiago, Chile

Abstract

The publication of semantic web data, commonly represented in RDF, has experienced outstanding growth over the last few years. Data from all fields of knowledge are shared publicly and interconnected in active initiatives such as Linked Open Data. However, despite the increasing availability of applications managing large-scale RDF information such as RDF stores and reasoning tools, little attention has been given to the structural features emerging in real-world RDF data. Our work addresses this issue by proposing specific metrics to characterize RDF data. We specifically focus on revealing the redundancy of each dataset, as well as common structural patterns. We evaluate the proposed metrics on several datasets, which cover a wide range of designs and models. Our findings provide a basis for more efficient RDF data structures, indexes and compressors.

Keywords

RDF structure, RDF metrics, RDF features, Linked Data

1. Introduction

The Linked Data paradigm [1] converts raw data into first class citizens of the Web. It materializes the Semantic Web foundations and enables raw data from diverse fields to be interconnected within data-to-data cloud. The Resource Description Framework (RDF) [2] is the increasingly cornerstone of this semantic approach. RDF provides a graph-based data model to structure and link data that describe things in the world. Its semantic model is extremely simple; a description of an entity (also called resource) is represented through triples in the form *(subject, predicate, object)*. Thus, a dataset in RDF data can be seen as a graph of knowledge in which subject entities and object values are linked via labeled edges.

The use of RDF to expose semantic data has seen a dramatic increase over the last years, making RDF data ubiquitous. As an example, LODStats[1], a project constantly monitoring the Linked Open Data cloud[2], reports (in May 2016) 2,832 live datasets having 150 billion triples. Part of this success of RDF is due to the graph conception and its expressive, but flexible, power: conceived as a semi-structured, open-world philosophy, the labeled graph structure underlying to the RDF model enables to add new properties and entity descriptions on demand.

Efficient RDF indexing [3,4], RDF compression [5,6] or distributed RDF management [7,8], to mention but a few, are novel areas emerging to cope with the scalability challenges of large-scale RDF processing. While general-purpose graph-based tools may be adequate for managing RDF, most approaches focus on providing native solutions tailored to RDF, hence they can take advantage of its particularities to boost performance.

Corresponding author:
Javier D. Fernández, Vienna University of Economics and Business, Institute for Information Business, Welthandelsplatz 1, Building D2, Entrance C, 2nd Floor, 1020 Vienna (Austria)
Email: javier.fernandez@wu.ac.at

In this scenario, there is a growing need for characterizing structural properties of real-world RDF data, which, however, is not completely covered in the state of the art. In this regard, initial works inspected the presence of power-law distributions [9,10], or study network-based features, such as clustering coefficient and path lengths [11,12]. Few studies move away from this line of research and further inspect low-level RDF particularities. For instance, what is the frequency of multivalued pairs (e.g. subject, predicate)? How many subjects act also as objects in other relations? Do typed subjects (subject with a defined rdf:type) present different features? To the best of our knowledge, this fine-grained analysis of individual RDF datasets has not been addressed systematically.

In this paper we present a theoretical and empirical study on real-world RDF structure and properties, in order to determine common features and characterize real-world RDF data. Our purpose is not to serve as a one-size-fits-all set of metrics, but to complement state-of-the-art graph-based features and provide a handbook with simple but useful metrics when developing efficient RDF data structures, indexes and compression techniques.

The rest of the paper is organized as follows. Section 2 presents the sparingly number of studies addressing real-world RDF structural characterization. We present our metrics for RDF graphs in Section 3, providing theoretical foundation as well as practical motivation and application for each proposed metric. In Section 4 we provide a fine-grained analysis our metrics on a real-world corpus of RDF datasets. Finally, we summarize and discuss the results in Section 5, pointing out applications and future work.

2. Related Work

The study of the RDF structure traditionally leads with two important and correlated aspects, *part-whole* and *schema-instance* dualities. On the one hand, part-whole distinguishes between the study of the structure of a single RDF dataset and the consideration of the whole Linked Open Data as a network of networks [13]. On the other hand, schema-instance considers that the semantics of RDF can be completed with (lightweight) ontologies defining a schema of the data (e.g. by means of the built-in vocabulary provided by RDFS [14] and OWL [15]), hence a structural characterization can whether study the ontology structure independently or to consider it implicitly in the data. In the next sections, we will propose and evaluate a set of metrics focused on a single RDF dataset at the instance level.

Power Law Distributions/Scale-free Network. RDF data are not random graphs [16] where, at large scale, the probability that a vertex has certain degree *k* follows a Poisson distribution. One of the first conclusions of initial RDF studies was that RDF graphs, instead, follow power law distributions in most of their metrics [9]. A power law is a function with scale invariance (scale-free), which can be drawn as a line in the log-log scale with a slope equal to a scaling exponent. Empirical observations in real networks have found fat-tailed and scale-free structures in several real-world graphs such as the WWW [17], scientific citation nets [18] and protein-protein interactions [19].

In RDF graphs, Ding and Finin [9] crawled more than 300 million triples from 1.7 million documents finding power law distributions in metrics such as (i) the number of RDF documents per website, (ii) the number of triples per RDF document, and (iii) the use of instances of the defined classes and properties, reporting that 97% of classes and 70% of properties are defined but never used. They also showed that most resources are described with two to ten triples.

Bachlechner and Strang [11] collected more than 1.6 million Fiend-Of-a-Friend (FOAF[3]) documents, reaching similar conclusions for the in- and out-degree distributions (number of triples related to a subject, and the number of triples related to an object, respectively) in each community, as well as the entire network. They reported an average degree of 9.56, whereas the maximum was 7,739, reflecting its skewed distribution.

Ge et al [12] define the notion of Object Link Graph, considering an undirected graph of related resources. An empirical study on 110.5M Web crawled resources as well as individual datasets (such as DBpedia[4] and BIO2RDF[5]) revealed that the Object Link Graph also holds a power law distribution.

Focused on the schema level, Theoharis et al [10] studied 250 Semantic Web schemas (RDFS and OWL), finding power law distributions in 58.6% of the schemas, in total-degrees (sum of in- and out- degree), out- and in-degrees. Zhang [20], on two biomedical ontologies, and Hu et al [21], on 4,433 ontologies, also confirm this distribution.

Small-world Phenomenon. A graph is actually a small world when it has short global separations, i.e., the average minimum distance between nodes is limited [22]. It is also associated with high local clustering (bigger than a random graph). The small-world phenomenon has been popularly accepted within the networks of friends, stating that two random citizens are connected by only six degrees (intermediate nodes) of difference [23]. In practice, small-world networks have several important characteristics, such as large presence of cliques (subgraphs in which all the possible connections are present) and hubs (intermediate nodes with many associations, i.e., high degree). These latter are used to navigate through the network in few steps, and are good candidates for feeding them as seeds in the search engine [12].

The consideration of the Semantic Web as a small world is mostly accepted. Bachlechner and Strang [11] evaluated FOAF communities, finding high clustering coefficients in all subgraphs. Gil and García [24] computed the 1-neighborhood clustering coefficient for a directed graph at the schema level, obtaining a slightly smaller clustering coefficient than the WWW factor of 0.108, as studied by Adamic [25]. Regarding the path lengths, Guns [26], with a small corpus of instances, established the longest shortest path (diameter) in 11 steps whereas the average was only 4.12. Note that the directed diameter of the Web is at least 28 (for the connected component [27]). Ge et al [12], with a bigger corpus, found an effective length of 11.53, almost the double than the 6.83 for the traditional WWW [27] when the direction of links is considered. In contrast, Gil and García [24] and Cheng and Qu [28], both at the schema level, found average path length of 5.07 and 10.05 respectively, denoting a high influence on the particular ontology design.

Other Studies. The presented studies mainly focus on analysing network-based metrics in RDF graphs. However, little attention has been given to low-level particularities of RDF, such as the repetitions of particular terms as well as pairs of elements (subject-predicate, subject-object, predicate-object) or the presence of frequent patterns when describing a subject. In this regard, cardinalities for (subject,predicate) and (predicate,object) pairs have been defined [29,30] with the aim of characterizing the data for particular purposes such as improving RDF compression techniques [6] or measuring the interlinkage and publication quality of RDF online [31]. Similar efforts are conducted by LOUPE [32], a tool to inspect online RDF datasets and show their main characteristics (e.g. types, structures, vocabularies, etc.). Our work aims at fulfilling this gap, extending these metrics to serve as a catalogue of low-level metrics tailored for RDF graphs.

3. Proposed Metrics for RDF Graphs

In the following, we provide specific metrics to characterize RDF data. We follow the standard RDF formalization [33,34]. Assume infinite, mutually disjoint sets U (RDF URI references), B (Blank nodes), and L (RDF literals).

Definition 1 (RDF triple). A tuple $(s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)$ is called an RDF triple, in which s is the subject, p the predicate and o the object.

Definition 2 (RDF graph). An RDF graph G is a set of RDF triples. Then, (s, p, o) can be represented as a direct edge-labeled graph $s \xrightarrow{p} o$.

For our purposes, we make no distinction between URIs, Blank nodes and Literals. We also note that the RDF interpretation as a graph can be misleading. As shown in Definition 1 and 2, an RDF dataset can be represented as an edge-labeled graph. This conception is useful for some purposes such as modelling or visualization. However, it cannot be considered as a labeled graph in the standard sense because the predicates can again appear as nodes of other edges [35], in order to define a schema over the data. Thus, the well-established methods from graph theory need to be slightly adapted to consider the seamless schema-instance integration provided in RDF.

3.1. Subject and Object Degrees

As stated, previous studies focused on showing the presence of power-law distributions on subject and object in- and out-degrees [9,11]. Although this is a useful indicator that some level of compression can be achieved [36], additional low-level details are needed to design more efficient RDF-based data structures (e.g. RDF compressors and indexes). To do so, we propose simple metrics on the characterization of such degrees. For the sake of clarity, we first summarize the purpose of each category prior to the formal definition:

out- and in- degrees: to known the cardinality of subjects and objects. A subject with a high out-degree is a so-called "star" (a resource described in depth). An object with a high in-degree tends to be a repeated final value or a hub to further information.

partial out- and in- degrees: to describe the presence and cardinality of the multivalued pairs *(subject,predicate)* and (*object,predicate*). That is to say, they quantify the number of objects related to the same (*subject,predicate*) and the number of subjects for a given (*object,predicate*).

labeled out- and in- degrees: to know the number of different predicates related to subjects and objects. It shows if subjects are described with many predicates and, respectively, if object values are used with many predicates.

direct out- and in- degrees: to count direct relationships between subjects and objects, thus minimizing the effect of the labelling. They consider to disregard labels and to count the number of objects related to a subject and, respectively, the corresponding number of subjects for each object.

Let *G* be an RDF graph, and S_G, P_G, O_G be the sets of subjects, predicates and objects in *G*. Assume generic $s \in S_G$, $p \in P_G$ and $o \in O_G$. Let us also denote Z_G and X_G the set of valid pairs (*subject,predicate*) and (*object,predicate*) respectively. That is, $Z_G = \{(s,p) \mid \exists z: (s,p,z) \in G\}$, and $X_G = \{(o,p) \mid \exists\ x: (x,p,o) \in G\}$.

Definition 3 (out-degree). The **out-degree** of s, denoted $deg^-(s)$, is the number of triples in G in which s occurs as subject. Equation 1 provides its formal definition.

$$deg^-(s) = |\{(s,y,z)|(s,y,z) \in G\}| \quad \textbf{(1)}$$

In turn, we define the **maximum** and **mean** out-degrees of G, $deg^-(G)$ and $\overline{deg^-}(G)$ respectively, as the maximum and mean out-degrees of all subjects (S_G).

Definition 4 (partial out-degree) The **partial out-degree** of s with respect to p, denoted $deg^{--}(s,p)$, is defined as the number of triples of G in which s occurs as subject and p as predicate. Its formal definition is provided in Equation 2.

$$deg^{--}(s,p) = |\{(s,p,z)|(s,p,z) \in G\}| \quad \textbf{(2)}$$

While, we also define the **maximum** and the **mean** partial out-degree of G, $deg^{--}(G)$ and $\overline{deg^{--}}(G)$ respectively, as the maximum (resp. the mean) partial out-degrees of all pairs of subject-predicates (Z_G).

Definition 5 (labeled out-degree) The **labeled out-degree** of s, $deg_L^-(s)$, is defined as the number of different predicates (labels) of G with which s is related as a subject in a triple of G. Equation 3 provides its formal definition.

$$deg_L^-(s) = |\{p|\exists z \in O_G, (s,p,z) \in G\}| \quad \textbf{(3)}$$

In turn, we define the **maximum and mean** labeled out-degree**,** $deg_L^-(G)$ and $\overline{deg_L^-}(G)$ respectively, as the maximum (resp. the mean) labeled out-degrees of all subjects (S_G).

Definition 6 (direct out-degree*)* The **direct out-degree** of s, denoted $deg_D^-(s)$, is defined as the number of different objects of G with which s is related as a subject in a triple of graph G. Its formal definition is provided in Equation 4.

$$deg_D^-(s) = |\{o|\exists y \in P_G, (s,y,o) \in G\}| \quad \textbf{(4)}$$

We also define the **maximum** and **mean** direct out-degrees, $deg_D^-(G)$ and $\overline{deg_D^-}(G)$ respectively, as the maximum (resp. the mean) direct out-degrees of all subjects of G. It is worth noting that, given the definition, the *direct out-degree* of a subject *s* can only differ from its *out-degree* when *s* is related to, at least, an object *o* by means of two or more different predicates. In other words, if every (*subject,object*) pair is only related with one predicate, then *out-degrees* are equal to *direct out-degrees*.

Symmetrically, we define the *in-degrees* for objects in a formal way (for the sake of simplicity, we omit the maximum and mean in-degrees, which can be defined similarly to the in-degrees).

Definition 7 (in-degree) The **in-degree** of o, denoted $deg^+(o)$, is the number of triples in G in which o occurs as object.

$$deg^+(o) = |\{(x,y,o)|(x,y,o) \in G\}| \quad \textbf{(5)}$$

Definition 8 (partial in-degree) The partial in-degree of o with respect to p, denoted $deg^{++}(o,p)$, is defined as the number of triples of G in which o occurs as object and p as a predicate.

$$deg^{++}(o,p) = |\{(x,p,o)|(x,p,o) \in G\}| \quad \textbf{(6)}$$

Definition 9 (labeled in-degree) The **labeled in-degree** of o, denoted $deg_L^+(o)$, is defined as the number of different predicates (labels) of G with which o is related as object in a triple of G.

$$deg_L^+(o) = |\{p|\exists x \in S_G, (x,p,o) \in G\}| \quad \textbf{(7)}$$

Definition 10 (direct in-degree) The **direct in-degree** of o, denoted $deg_D^+(o)$, is defined as the number of different subjects of G with which o is related as an object in a triple of graph G.

$$deg_D^+(o) = |\{s|\exists y \in P_G, (s,y,o) \in G\}| \quad \textbf{(8)}$$

Note that the *cardinality*, *average cardinality*, *inverse cardinality* and *average inverse cardinality* metrics by Hogan et al [29] are equivalent to partial out-degree, average partial out-degree, partial in-degree and average partial in-degree.

3.1.1. Example and potential uses

We illustrate these properties in a small example graph presented in Figure 1. The graph models a resource *John*, who is an instance of the class *Researcher* and has two mail boxes. We also represent his birth place, named "*Roma*" in Italian, and two resources, *Giacomo* and *Piero*, whose area of work is also *Roma*.

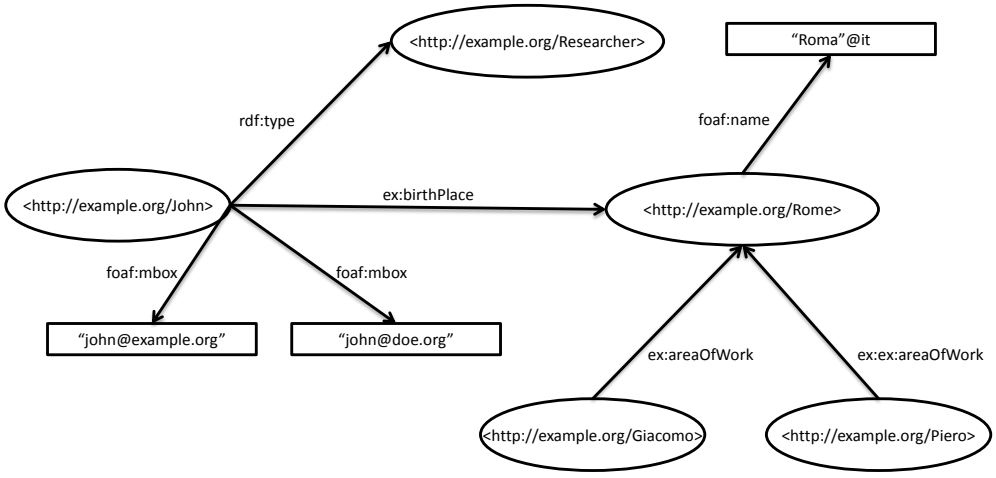

Figure 1. Running example: A basic RDF graph.

Table 1 reports the metrics for this example. In the example, the node *http://example.org/John* has a significant out-degree (it is related to four nodes, above average) and hence it conforms a star-shaped node. When designing an RDF data structure, *e.g.* an index, it is potentially interesting to know the presence or absence of these nodes, but also the distribution of these high out-degrees. For instance, if a real-world RDF graph has a maximum out-degree close to 1, it stands for a very simple graph whose access may be optimized. In contrast, a skewed distribution of out-degrees could require a more refined structure than the previous case. Thus, out-degree distribution together with maximum and mean values constitutes a fair characterization of these nodes in a given graph. Similar reasoning can be made for object in-degree, where the node is not a source, but is a common destination object node.

Table 1. Summary of structural metrics describing the running example.

Metric				Value	Metric				Value
SUBJECT OUT-DEGREE	Max	total	$deg^{-}(G)$	4.00	OBJECT IN-DEGREE	Max	total	$deg^{+}(G)$	3.00
		partial	$deg^{--}(G)$	2.00			partial	$deg^{++}(G)$	2.00
		labeled	$deg_L^{-}(G)$	3.00			labeled	$deg_L^{+}(G)$	2.00
		direct	$deg_D^{-}(G)$	4.00			direct	$deg_D^{+}(G)$	3.00
	Mean	total	$\overline{deg^{-}}(G)$	1.75		Mean	total	$\overline{deg^{+}}(G)$	1.40
		partial	$\overline{deg^{--}}(G)$	1.17			partial	$\overline{deg^{++}}(G)$	1.17
		labeled	$\overline{deg_L^{-}}(G)$	1.50			labeled	$\overline{deg_L^{+}}(G)$	1.20
		direct	$\overline{deg_D^{-}}(G)$	1.75			direct	$\overline{deg_D^{+}}(G)$	1.40
PREDICATE DEGREE	Max	total	$deg_P(G)$	2.00	RATIOS			$\propto_{s-o}(G)$	0.13
		out	$deg_P^{-}(G)$	2.00				$\propto_{s-p}(G)$	0.00
		In	$deg_P^{+}(G)$	2.00				$\propto_{p-o}(G)$	0.00
	Mean	total	$deg_P(G)$	1.40					
		out	$\overline{deg_P^{-}}(G)$	1.20					
		In	$\overline{deg_P^{+}}(G)$	1.20					

Regarding partial and labeled out- and in- degrees, they provide information on the different types of edges coming out from (or going into) a node. Partial degree provides a metric of the multi evaluation of pairs (subject-predicate or predicate-object), while labeled degree refines the nodes categorization. In the example, *http://example.org/Rome* is a common object as three subjects are related to it, hence its in-degree is three. However, the labeled in-degree is "two" as it receives edges from two labels *ex:birthPlace* and *ex:areaOfWork*. Subsequently, its partial in-degree is two, denoting that the pair (*http://example.org/Rome, ex:areaOfWork*) is multivalued.

As we shown in the forthcoming evaluation, labeled out-degree verifies that few predicates are related to the same subject or object. This could allow RDF structures for optimizing the representation of the list of predicates related to a given subject or object.

Finally, direct out- and in-degrees complete the degree metrics for subject and objects. They indicate the cardinality of binary relations between subjects and objects disregarding the labels. In the example, direct degrees throw similar results as the out- and in-degrees, as every *(subject,object)* pair is related only with one predicate. Direct degrees are useful when representing RDF as a classical adjacency matrix, e.g. representing subjects in rows and objects in columns. In such scenario, direct out-degrees model the cardinality of rows, whereas direct in-degrees describe the cardinality in columns.

3.2. Predicate Degrees

Despite the fact that important RDF characteristics can be extracted from the previous metrics (or a combination of them), one could argue that some RDF indexing techniques need further details. For instance, the family of indexing techniques following vertical partitioning [37] builds indexes per predicate. Typically, these techniques index all the *(subject,object)* pairs for each predicate. In such scenario, the number of *(subject,object)* pairs for each predicate would be a good indicator of the size and distribution of these predicate partitions.

With this objective in mind, we detail predicate degrees following the same preceding principles of simplicity and use in other scenarios. The purpose of the metrics is summarized as follows:

- **Predicate degrees:** to know the cardinality of predicates. In contrast to the relational model in which every row of a table is described with the same number of attributes (columns), the flexibility of RDF yields to a potentially high variability in the number of predicates describing each subject. Thus, this metric is an important clue to find the most used predicates in a given RDF dataset.
- **Predicate in- degrees:** to describe the number of subjects related to given predicates. It serves to refine knowledge about the distribution of subjects per predicate.
- **Predicate out- degrees:** to know the number of different objects related to given predicates, also used to describe the predicate degree in detail.

We make use of the aforementioned notation, being *G* an RDF graph, with S_G, P_G, O_G the sets of subjects, predicates and objects in *G* and generic $s \in S_G$, $p \in P_G$ and $o \in O_G$.

Definition 11 (predicate degree) The **predicate degree** of p, denoted $deg_P(p)$, is defined as the number of triples of graph G in which p occurs as predicate.

$$deg_P(p) = |\{(x,p,z)|(x,p,z) \in G\}| \quad \textbf{(9)}$$

Definition 12 (predicate in-degree) The **predicate in-degree** of p, denoted $deg_P^+(p)$, is defined as the number of different subjects of G with which p is related as a predicate in a triple of G.

$$deg_P^+(p) = |\{s|\exists z \in O_G, (s,p,z) \in G\}| \quad \textbf{(10)}$$

Definition 13 (predicate out-degree) The **predicate out-degree** of p, denoted $deg_P^-(p)$, is defined as the number of different objects of G with which p is related as a predicate in a triple of G.

$$deg_P^-(p) = |\{o|\exists x \in S_G, (x,p,o) \in G\}| \quad \textbf{(11)}$$

Analogously, the **maximum** and **mean predicate degree, in-degree and out-degree** are defined as the maximum and mean predicate corresponding degrees of all predicates in G.

3.2.1. Explanation and potential uses.

As stated, the predicate degree constitutes an essential metric when a *(subject,object)* or *(object,subject)* is built for each predicate, such as the vertical partitioning technique [37].

The predicate degree reflects the number of entries for a predicate partitioning. In turn, the predicate in-degree and out-degree refine this metric by providing the domain and range sizes for each predicate. For instance, predicates such as *rdf:type* have a limited range (low predicate out-degree) but a great domain (high predicate in-degree). In turn, if a predicate returns a high degree (it appears in many triples) but a low out-degree, it reveals that few values are repeated along descriptions. For instance, this is the case of discrete values for predicates such as "City_State" or "Postal_code" in which a dozen of similar values could be repeated in thousands or millions of records.

Figure 1 illustrates these metrics. Despite the limited size of the example, it shows the variable figures of predicate degrees. For instance, *foaf:name* is present only once whereas *foaf:mbox* and *ex:areaOfWork* are twice. In this latter, its predicate in-degree is two (denoting two different subjects) yet the out-degree is only one (two subjects point to the same object). This shows that predicate in- and out- degree could roughly classify predicate usage as follows:

- **N:N predicates,** having a similar in- and out-degree, i.e., $\deg_P^+(p) \cong \deg_P^-(p)$. Note that a special case would be 1:1 predicates, i.e. predicates appearing only in one triple, but this is a marginal case at large scale. In general, it is accepted that the number of predicates is much smaller than the number of subjects and objects [38].
- **1:N predicates,** having a significant smaller in-degree than their out-degree, $\deg_P^+(p) \ll \deg_P^-(p)$.
- **N:1 predicates,** having a significant greater in-degree than their out-degree, $\deg_P^+(p) \gg \deg_P^-(p)$.

Although the formal demonstration of this classification goes beyond the purpose of this paper, one could envision that this is a general scenario in real-world datasets. For instance, predicates describing unique IDs, such as "Passport" or "Protein_ID", belong to *1:1 predicates*. In turn, the mentioned "City_State" or "Postal_code" fall into *N:1 predicates*. Finally, other predicates, such as "foaf:mbox" in the example, can belong to *1:N predicates*.

3.3. Common Ratios

The presence of star nodes is popularly accepted as a natural consequence when describing a resource in depth. A second popular "construction" is the presence of chains, *i.e.*, paths of linked nodes. This construction occurs, for instance, whenever we use *owl:sameAs* to interlink two described entities. As some of these nodes in the chain is also a star, one could talk of "star chained design" for RDF datasets.

Intermediate nodes in chains appear in two triples acting with different roles. For instance, let us suppose a design such as $A \overset{p_1}{\rightarrow} B$ and $B \overset{p_2}{\rightarrow} C$. As shown, B is present in two triples, being an object in the first one, and subject in the latter. Additionally, we should also consider that predicates can again appear as nodes of other edges, acting also as

intermediate nodes. In general terms, considering the three different roles in triples (subjects, predicates and objects), there could exist elements which are present in a graph acting with more than one role.

We use three metrics to characterize the proportion of common elements with respect to the total elements. In short:

- **Subject-object ratio**: to describe the number of elements acting both as subject and objects among all subjects and objects. In other words, the subject-object ratio denotes the percentage of nodes having incoming and outgoing edges. They are, in fact, the main players when navigating the graph.
- **Subject-predicate ratio**: to describe the number of elements acting both as subject and predicates among all subjects and predicates. This points that semantics is given to predicates, *e.g.* using *rdfs:domain* or *rdfs:range*.
- **Predicate-object ratio**: to describe the number of elements acting both as predicates and objects among all predicates and objects. It refines the previous metrics, *e.g.* when using *rdfs:subPropertyOf*.

Formally described, let us retake again G as an RDF graph, with S_G, P_G, O_G the sets of subjects.

Definition 14 (subject-object ratio $\alpha_{s\text{-}o}$) The subject-object ratio $\alpha_{s\text{-}o}(G)$ of a graph G is defined as the ratio of common subjects and objects in the graph G.

$$\propto_{s-o} (G) = \frac{|S_G \cap O_G|}{|S_G \cup O_G|} \quad \textbf{(12)}$$

Definition 15 (subject-predicate ratio $\alpha_{s\text{-}p}$) The subject-predicate ratio $\alpha_{s\text{-}p}(G)$ of a graph G is defined as the ratio of common subjects and predicates in the graph G.

$$\propto_{s-p} (G) = \frac{|S_G \cap P_G|}{|S_G \cup P_G|} \quad \textbf{(13)}$$

Definition 16 (predicate-object ratio $\alpha_{p\text{-}o}$) The predicate-object ratio $\alpha_{p\text{-}o}(G)$ of a graph G is defined as the ratio of common predicates and objects in the graph G.

$$\propto_{p-o} (G) = \frac{|P_G \cap O_G|}{|P_G \cup O_G|} \quad \textbf{(14)}$$

3.3.1. Explanation and potential uses.

Ratios give evidence of chain constructions. Figure 1 illustrates that there are no common subject-predicates and predicates-objects. In contrast, in the previous example, the subject-object ratio reveals that 13% of the subjects and objects are common elements which take part in a subject-object path.

Subject-object is, in fact, the most common construction as it is a natural way of linking the description of two resources. Thus, this ratio provides a good measure of potential paths and the level of "navigability".

In turn subject-predicate and predicate-object ratios, when present, show how far predicates are also used as subjects or objects. These two ratios can be used to justify the consideration (or not) of a given RDF dataset as a graph. If there is a null influence of these types of shared nodes, one could assume that little semantics has been added.

3.4. Subject-Object Degrees.

Given the importance of subject-object nodes, a fine-grained analysis can be made. In particular, one could study the in- and out-degrees restricted to subject-object nodes. We define these degrees implicitly, as their formalization is equivalent to the degrees presented in Section 3.1, but restricted to subject-object nodes. For instance, the maximum out-degree of the graph G restricted to subject-objects, which is denoted as $deg^-(G)|_{s-o}$ is the maximum out-degree of all subject-object nodes in the graph G. Its formal definition is provided in Equation 15, while Equation 16 defines the mean out-degree of the graph G restricted to subject-objects.

$$deg^-(G)|_{s-o} = max_{s \in S_G \cap O_G}(deg^-(s)) \quad \textbf{(15)}$$

$$\overline{deg^-}(G)|_{s-o} = \frac{1}{|S_G \cap O_G|} \sum_{s \in S_G \cap O_G}(deg^-(s)) \quad \textbf{(16)}$$

3.4.1. Explanation and potential uses.

These metrics serve the same purposes as the original ones in Section 3.1, but restricted to subject-object nodes. This enables to provide a more detailed vision of what is going on in these important, intermediate nodes.

This characterization might result especially useful when common subject-objects connect two different graphs. In this case, one could grasp the features of these "connecting nodes" with these metrics, gaining insights to improve navigability. For instance, additional structures and indexes can be built for query suggestion or visualization purposes.

3.5. Predicate Lists.

The list of predicates related to a subject may vary greatly for each subject. However, there would exist repetitions whenever two subjects are described in the same way. For instance, the list of predicates used to describe a *song* varies enormously from those used to categorize a *protein*, and both can coexist in a cross-domain dataset. We define metrics to characterize these lists. In short:

- **Number and ratio of predicate lists:** it counts the different lists, and the ratio of lists from the total lists.
- **Degree of predicate lists:** it characterizes the number of repetitions of each list.
- **Lists per predicate:** it counts the number of different lists including each predicate.

Formally described, let Ls be the set of predicates (labels) related to the subject s. That is, the set of predicates Ls = $\{p \mid \exists z \in O_G, (s, p, z) \in G\}$. We denote as L_G to the set of different predicate lists in G. That is, $L_G = \{L_x, x \in S_G\}$, hence the number of different lists in the graph G is $|L_G|$. Note that the total predicate lists (with repetitions) is equal to the number of different subjects S_G.

Definition 17 (Ratio of repeated predicate lists) The ratio of repeated predicate lists $r_L(G)$ of a graph G is defined as the ratio of repeated predicate lists from the total lists in the graph G.

$$r_L(G) = 1 - \frac{L_G}{S_G} \quad \textbf{(17)}$$

Definition 18 (predicate list degree) The predicate list degree of a predicate list Ls, denoted $deg_{PL}(L_S)$, is defined as the number of different subjects in G whose list of predicates is exactly L_s. Equation 18 provides its formal definition. In turn, Equations 19 and 20 defines the maximum predicate list degree, and resp. the mean predicate list degree of the graph G, as the maximum and mean out-degrees of all predicate lists in G.

$$deg_{PL}(L_S) = |\{L_x | \, x \in S_G, L_x = L_s\}| \quad \textbf{(18)}$$

$$deg_{PL}(G) = max_{L_x \in L_G}\,(deg_{PL}(L_x)) \quad \textbf{(19)}$$

$$\overline{deg_{PL}}(G) = \frac{1}{|L_G|}\sum_{L_x \in L_G} deg_{PL}(L_x) \quad \textbf{(20)}$$

Definition 19 (lists per predicate degree) The lists per predicate degree of a predicate p, $deg_{LPP}(p)$, is defined as the number of different predicate lists in L_G in which the predicate appears.

$$deg_{LPP}(L_S) = |\{L_x | \, p \in S_G, L_x = L_G\}| \quad \textbf{(21)}$$

The maximum and resp. the mean lists per predicate degree of the graph can be defined as the maximum and mean out-degrees of all predicates in G.

3.5.1. Explanation and potential uses.

The metrics for the predicate lists characterize the repetition of predicate structures. On the one hand, if a short set of lists is present (highly structured data), one could perfectly categorize and manage a reduce set of combinations. On the other hand, "random" lists denote the presence of a cross-domain datasets or a light schema, as few repetitions are present.

The example in Figure 1 presents four predicate lists (one per subject): [*rdf:type, ex:birthPlace, foaf:mbox*], [*foaf:name*], and [*ex:areaOfWork*], repeated twice. This repetition denotes a common pattern in the data (despite the

reduced size of the example). In fact, the ratio of repeated predicate lists is $r_L(G) = 1 - 3/4 = 0.25$. This means that 25% of the predicate lists are repetitions. Note also that each predicate is present in only one list. In other words, in this particular case, predicates are unequivocally included in one list.

Predicate lists characterization would serve several purposes. For instance, for large-scale visualization, Khatchadourian and Consens [39] group common predicate structures to summarize the links between Linked Open datasets, hence these metrics may contribute by categorizing the type of repetitions. In turn, several RDF indexing approaches consider the commonalities in the predicate structures. Campinas et al [40] make a structural summary grouping the entities having the same set of predicates in order to suggest potential predicates and relationships when writing a query. Tran et al [41] propose a structure index for RDF, grouping similar structured data elements. Hernández-Illera et al [42] follow a similar approach as a preprocessing step to compress RDF data. In these cases, the proposed metrics may help in determining structural properties of the indexes.

3.6. Typed Subjects and Classes

RDF resources can be associated to types by means of the *rdf:type* predicate. The values for this predicate are then *Classes*, which can be described in detail by means of RDFS [14]. For instance, in the example in Figure 1, *John* is of type *Researcher*. One should expect that, as previously mentioned, entities of the same class would be described with similar predicates. We define metrics to characterize these commonalities. In short:

- **Number of classes**: it counts the number of different classes.
- **Number and ratio of typed subject**: it counts the number of typed subjects (those including at least one type) and the ratio over the total subjects.
- **Lists per class**: it counts the number of different predicate lists included in each class.
- **Out-degrees of typed subject**: it characterizes the out-degrees of typed subjects.
- **Degree of predicate lists for typed subjects**: it characterizes the number of repetitions of those predicate list including at least one *rdf:type*.

Formally, let C_G be the set of all classes in the graph G, and c a generic class, $c \in C_G$. The **number of all different classes** is then $|C_G|$. Let S^c be the set of subjects of type c, $S^C = \{s|(s,t,c) \in G\}$, being t the predicate *rdf:type*. The set S_G^C denotes all different typed subjects in the graph G, that is $S_G^C = \{s|\exists c \in C_G, (s,t,c) \in G\}$, with predicate t= *rdf:type*. The number of different typed subjects in the graph is then $|S_G^C|$.

Definition 20 (Ratio of typed subjects). The ratio of typed subjects $r_T(G)$ of a graph G is defined as the ratio of different typed subjects from the total subjects of G.

$$r_T(G) = \frac{|S_G^C|}{|S_G|} \quad \textbf{(22)}$$

Let L_G^C be the set of different predicate lists for typed subjects. That is, formally described, $L_G^C = \{L_x, x \in S_G^C\}$.

Definition 21 (lists per class degree). The lists per class degree} of a class c, $deg_{LPC}(c)$, is defined as the number of different predicate lists in L_G in which the class c appears as a value for a typed subject.

$$deg_{LPC}(c) = |\{L_x|L_x \in L_G^C, x \in S^C\}| \quad \textbf{(23)}$$

The maximum and resp. the mean lists per class degree of the graph can be defined as the maximum and mean out-degrees of all classes in G.

We define the *typed subject out-degrees* and the *degree of predicate lists for typed subjects* implicitly, as their formalization is straightforward. In the first case, the *typed subject out-degrees* are equivalent to those studied in Section 3.1. but restricted to typed subjects. For instance, the *maximum out-degree* of the graph G restricted to typed subjects, which is denoted as $deg^-(G)|_{S_G^C}$ is the maximum out-degree of all typed subjects in the graph G. Its formal definition is provided in Equation 24, while Equation 25 defines the mean out-degree of the graph G restricted to typed subjects.

$$deg^-(G)|_{S_G^C} = max_{s \in S_G^C}(deg^-(s)) \quad \textbf{(24)}$$

$$\overline{deg^-}(G)|_{S_G^C} = \frac{1}{|S_G^C|}\sum_{s \in S_G^C} deg^-(s) \quad \textbf{(25)}$$

3.6.1. Explanation and potential uses.

The characterization of different classes and typed subjects, as well as their degrees, is an important step in describing a common schema, if present. As we have motivated, one should expect that subjects typed equally would be described with similar predicates. These metrics provide an answer to this assumption, and give insights of other schema features. For instance, the ratio of typed subjects constitutes a ratio of the level of well-categorized information. They also help determine if typed subjects are (or not) further described than non-typed ones.

In our example in Figure 1 only one class (*Researcher*) and one typed subject (*John*) are present. As there are four different subjects, the ratio of typed subjects is 0.25. In this simple example, there is only one predicate list per class, [rdf:type, ex:birthPlace, foaf:mbox]. As for the previous predicate list metrics, this characterization may serve diverse purposes, e.g. visualization [40] and structural indexing [41,42], but also reasoning. For this latter, we characterize not only the presence of instances for the classes, but the different predicate lists, which may be useful to create a reduced index with all the possible variants.

4. Experimental Evaluation

We perform an evaluation to illustrate the proposed metrics on real-world RDF datasets. Thus, we first establish an experimental framework (Section 4.1) and we compute and present the aforementioned parameters (Section 4.2). Note that the results on this corpus should not be extrapolated to the whole linked open data cloud, but they show the use of the metrics to characterize an RDF dataset and gain insights toward the aforementioned potential uses. For a comprehensive explanation, the order of presentation of the results is slightly different than the previous definitions.

4.1. Experimental Framework

Table 2 shows our experimental framework. We define **seven categories**: media, publications, knowledge base, government, sensors, geography and biology. This distinction is based on the Linked Open Data cloud topics. We choose fourteen datasets based on the number of triples, topic coverage, availability and previous uses in benchmarking. Most datasets are well-known in the area:

- **Media**: *Jamendo* represents music records and artists, *LinkedMDB* stores movies and authors, *Dbtune* provides music-related structured data, and *Flickr Event Media* (Flickr hereinafter) holds Flickr events and their authors.
- **Publication**: *SWDF* is a small dataset with information related to the main conferences and workshops in the area of Semantic Web, whereas *Faceted DBLP* (referred to as DBLP hereinafter) is an RDF conversion of the well-known bibliographic repository.
- **Knowledge Bases**: *Wordnet 3.0* is a conversion to RDF of Wordnet (a lexical database of English) and *Dbpedia 3-8* is an RDF conversion of Wikipedia.
- **Government**: The *2011 Australian Census* is an open portion of the given census with aggregated data and the 2000 US Census comprises the first entities of the given census.
- **Sensors**: *AEMET* includes measurements made by the network of meteorological stations of the Spanish Meteorological Agency, and *Ike* contains meteorological sensor information of the real Ike hurricane.
- **Geography**: *Linked Geo Data* holds geographic information mainly from OpenStreetMap.
- **Biology**: *Affymetrix* contains probesets used in DNA microarrays.

A preprocessing phase is applied (all final datasets are publicly available[7]). First, for a fair comparison, we convert all datasets to N-Triples [43] (by means of the Any23 tool[6]), a basic format containing one sentence per line. Then, datasets are lexicographically sorted and duplicate triples are discarded. Table 2 reflects the number of triples, the final dataset size, and the number of different subjects, predicates, objects, and common subject-objects. As expected, the number of predicates remains commonly low. There are two exceptions: *Dbpedia* and *Linked Geo Data* are extreme cases in which the number of predicates grows to the order of thousands due to the variability of the represented information. However, note that the number of predicates remains proportionally small to the total number of triples.

Table 2. *Description of the evaluation framework.*

	Dataset	Triples	Size(MB)	Subjects	Predicates	Objects	Common SO
Media	Jamendo	1,049,637	144	335,925	26	440,602	290,291
	LinkedMDB	6,147,996	850	694,400	222	2,052,959	416,664
	Dbtune	58,920,361	9,566	12,401,228	394	14,264,221	10,076,199
	Flickr Event Media	49,107,168	6,714	5,490,007	23	15,041,664	3,822,727
Publications	SWDF	101,321	16	10,476	132	34,609	10,374
	Faceted DBLP	60,139,734	9,799	3,591,091	27	25,154,979	1,323,104
Knowledge	Wordnet 3.0	6,257,922	974	1,100,503	85	1,689,363	1,021,222
	Dbpedia 3-8	431,440,396	63,053	24,791,728	57,986	108,927,201	22,762,644
Government	2011 Australian Census	361,842	52	51,768	26	6,901	508
	2000 US Census	149,182,415	21,796	23,904,658	429	23,996,813	23,815,829
Sensors	AEMET	3,547,154	726	394,289	23	793,664	433
	Ike	514,824,008	102,662	114,484,017	10	114,629,189	114,484,017
Geography	Linked Geo Data	274,668,813	39,423	51,916,995	18,272	121,749,861	41,471,798
Biology	Affymetrix	44,207,145	6,526	1,421,763	105	13,240,270	245

4.2. Ratios

Table 3 shows the ratios (see Section 3.3.) of the evaluated datasets. They are a good starting point as they can reveal a level of cohesion between the different types of nodes and denote the presence (or absence) of shared nodes and labels.

Table 3 *Ratios of the given datasets.*

	Dataset	Common SO ($\propto_{s-o}(G)$)	Common SP ($\propto_{s-p}(G)$)	Common PO ($\propto_{p-o}(G)$)
Media	Jamendo	0.60	0	0
	LinkedMDB	0.18	0	$1.66*10^{-5}$
	Dbtune	0.61	0	$7344*10^{-6}$
	Flickr Event Media	0.23	0	0
Publications	SWDF	0.30	0	0
	Faceted DBLP	0.05	$7.52*10^{-6}$	0
Knowledge	Wordnet 3.0	0.58	$7.27*10^{-6}$	$1.78*10^{-6}$
	Dbpedia 3-8	0.21	$2.24*10^{-3}$	$7.50*10^{-5}$
Government	2011 Australian Census	0.01	$9.65*10^{-5}$	$8.67*10^{-4}$
	2000 US Census	0.99	0	0
Sensors	AEMET	$3.65*10^{-4}$	0	0
	Ike	0.99	0	0
Geography	Linked Geo Data	0.31	0	$4.52*10^{-7}$
Biology	Affymetrix	$1.67*10^{-5}$	0	$5.89*10^{-6}$

As expected, subject-object is the most frequent path constructor indeed and subject-predicate and predicate-object ratios are almost negligible. In fact, these latter are scheme descriptions, which are rare due to the RDF itself is schema-relaxed and the vocabulary can evolve as needed on demand.

The subject-object ratio shows interesting variable figures, ranging between 0 to 99%. Extreme cases are particularly of interest. For instance, the 2011 *Australian Census* and *AEMET* present values near to 0 whereas their counterparts per

category, the *2000 US Census* and *Ike* show values near of 99% of shared nodes. One can find the explanation in the diverse strategy followed to model the information. On the one hand, both the *2011 Australian Census* and *AEMET* describe particular values for a given entity (a statistic value or a sensor measure). Thus, a more "isolated" graph can be found in such cases where we represent certain measures. On the other hand, both the *2000 US Census* and *Ike* make use of intermediate nodes (blank nodes in the census and entity resources in Ike) to organize the different types of measures, resulting in a highly connected graph.

The low subject-object ratio in *Faceted DBLP* and *Affymetrix* is due to a different reason. In both cases, the datasets describe entities with a high number of different literals values. In the first case, titles, identifiers, dates, homepages, etc., of authors, articles and conferences are scarcely repeated. In the second, *Affymetrix* also describes entities (probesets) by different literal values (for labels, identifiers, version, description, dates, etc.). In addition, although URIs are used as objects, they are further described (as subjects) in other different datasets in the *bio2rdf* project.

The rest of the datasets can be grouped into two categories: datasets holding around 20-30% of shared entities (*LinkedMDB, Flickr, SWDF, Dbpedia* and *Linked Geo Data*), or near 60% (*Jamendo, Dbtune* and *Wordnet*).

4.3. Out- and in-degrees

In this section we study the mean out- and in-degree for subjects and objects respectively. The mean results and their standard deviations are presented in Figure 2. For the sake of comprehensibility, we erase hereinafter those error bars that significantly exceed the range of the figure. In this case, all in-degree deviations are erased. It is worth mentioning that both axes are in logarithmic scale. We also plot a dashed line delimiting the *1* value.

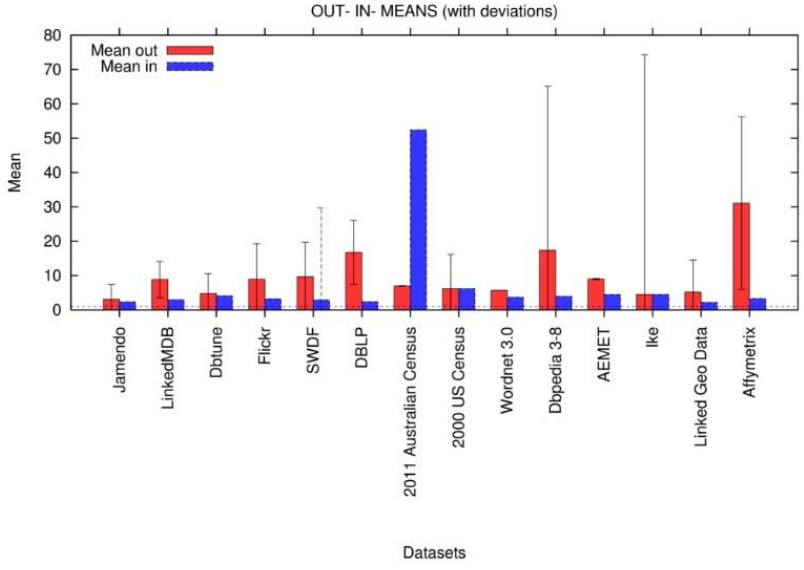

Figure 2 Mean out- and in-degrees for the evaluation datasets.

Most datasets present a limited mean number of triples per subject and object. Regarding the out-degree, its mean is modestly greater than 10 only for *DBLP*, *Dbpedia* and *Affymetrix*. This denotes that most datasets (even those with hundreds of millions of triples) present a mean of *10* triples per subject. In turn, the mean in-degree is even lower. All datasets have a lower mean in-degree than out-degree, being always smaller than *10*, i.e. objects appear in a mean of 10 triples. The exception is the *2011 Australian Census*, whose discrete object values are highly repeated in many triples.

Both mean out- and in-degrees show, in general, a high standard deviation. In fact, all in-degree deviations exceed considerably the range of the figure. This points to a noticeable skewed structure, more remarkable in objects.

Then, we study the out- and in-degrees, i.e. the subject and object distributions. Figure 3 illustrates these distributions for some representative datasets (*LinkedMDB*, *Dbpedia*, *AEMET* and *Ike*). In general, subjects and objects (out- and in-

degree) present skewed distributions. In fact, the in-degree in all datasets reveal a remarkably power law distributions in objects. Only *AEMET* (Figure 3, bottom left) slightly differs from the general tendency.

In turn, subjects (out degrees) also show clear power law distributions, represented in Figure 3 by *LinkedMDB* and *Dbpedia*. At this point, it is worth mentioning that the concrete power law exponent (denoting the slope of the line in a log-log scale) varies among datasets, ranging from -0.833 in *LinkedMDB* to -2.081 in *Dbpedia* (this latter is close to the -2.166 exponent by Ding and Finin [9]).

In contrast, a flat distribution is present in a reduced number of datasets, such as the two census datasets and AEMET (in Figure 3, bottom left). This reveals a data modelling where subjects are described with a similar number of triples. Finally, *Ike* distribution (Figure 3, bottom right) is a variation of the previous two types: some subjects are deeply described (or they have more relations) whereas others are concisely defined with few triples.

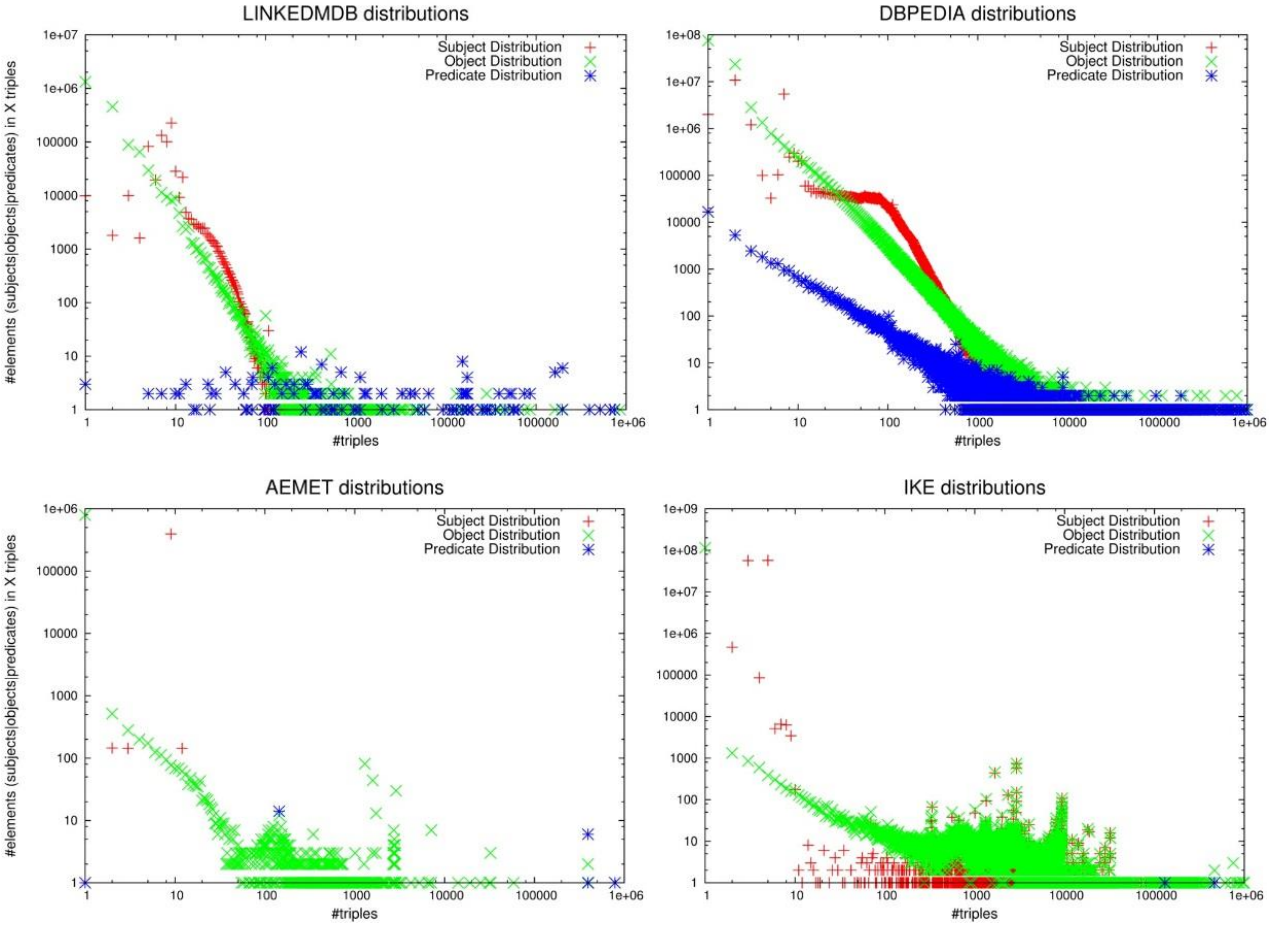

Figure 3 Degree distribution of representative datasets, in logarithmic scale.

Subject-object distribution. Figure 4 compares the previous mean out- and in-degree (presented in Figure 2) with the same degrees restricted to subject-object. For a fair comparison, we split the datasets by their range of common subject-object ratio (as stated in Section 4.2): common entities around 0%, 20-30%, 60% and 99%. We order the description of the results by these sets for explanation purposes:

- Common entities around 0%: In this case, the common entities are so rare that the means refer to few elements of the total. However, the mean in-degree restricted to these subject-objects is remarkably higher than for the total objects. We can find the reason of this difference in the non shared objects distribution. In all these datasets, a large number of different objects are present, whose in-degree is low (or even close to *1*) as we can

see in their corresponding in-degree distributions. Thus, the common subject-objects are more frequently present as they act as intermediate nodes and then playing as object in more triples on average.

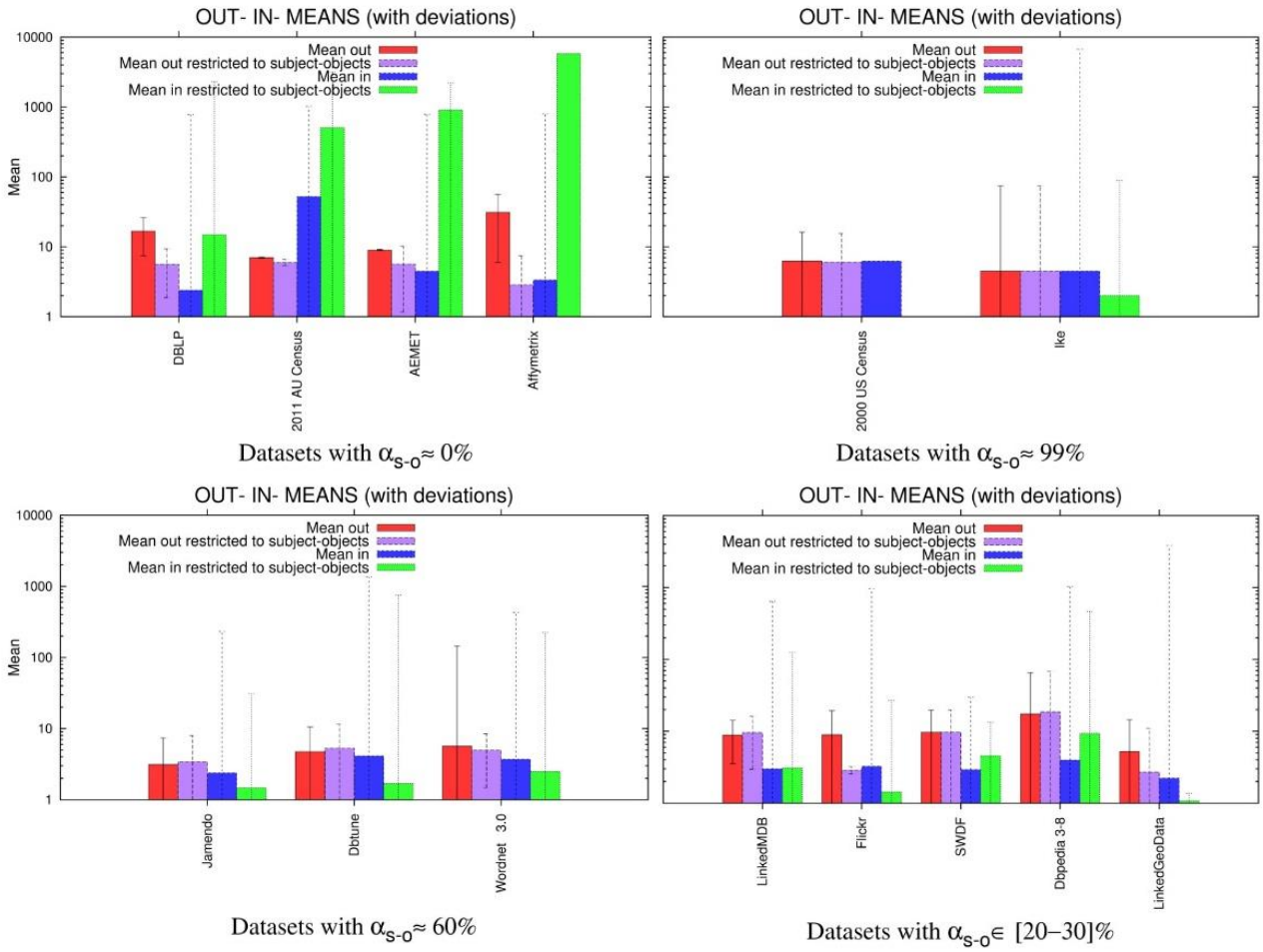

Figure 4 Mean out- and in-degrees for the evaluation datasets in comparison with the common subject-objects.

- Common entities around 99%: This is the case of the *2000 US Census* and *Ike*. Figure 4 shows low figures for the mean in-degree, being exactly *1* for the *2000 US Census*. We have argued that both datasets make use of different shared elements to organize the different types of figures or measures, hence the low in-degree. In contrast, given that 99% of elements are shared, the mean out-degree for these nodes is almost equal to the out-degree for all subjects.
- Common entities around 60%: The mean out-degrees are almost equivalent as more than 50% of the elements are shared, hence these nodes highly contribute to the original figures. As for the previous case of common entities around 99%, this scenario shows low figures for the mean in-degree. The reason in this case is equivalent as intermediate nodes organize the information.
- Common entities around 20-30%: This is the most variable set and datasets can present different results. In general terms, the mean out-degrees remain comparable. Nevertheless, *Flickr* and *Linked Geo Data* show a slightly smaller out-degree for subject-object nodes. This fact clearly depends on the represented information. For instance, this phenomenon can appear when an "event" in *Flickr* is described in depth but the related subject-object nodes representing "authors" are usually described in lesser depth. Regarding the in-degrees, in some cases the figures restricted to subject-objects are equal, slightly smaller or bigger than the non-restricted metric. The reasons are similar to the presented above: it would be slightly smaller for subject-objects when they serve to organize the information and slightly bigger whenever non repeated objects are predominant.

4.4. Predicates per Subject and Object

We study the labeled out- and in-degrees, that is, the predicates per subject and object respectively. Figure 5 illustrates the mean figures. As can be seen, the results show that few predicates are related to the same subject, on average. The only exception is *Affymetric*, were 20 predicates are present per subject. This fact, together with its large mean out-degree (more than 30 triples per subject) reflect a dataset design where entities are described in detail. In contrast, datasets such as *Jamendo* and *Ike* show a mean of 3-4 predicates per subject, reflecting a less diverse description. Note that, in all cases, the mean labeled out-degree is a clear indicator of the presence of star-shaped nodes, i.e., nodes with different triples around one common subject.

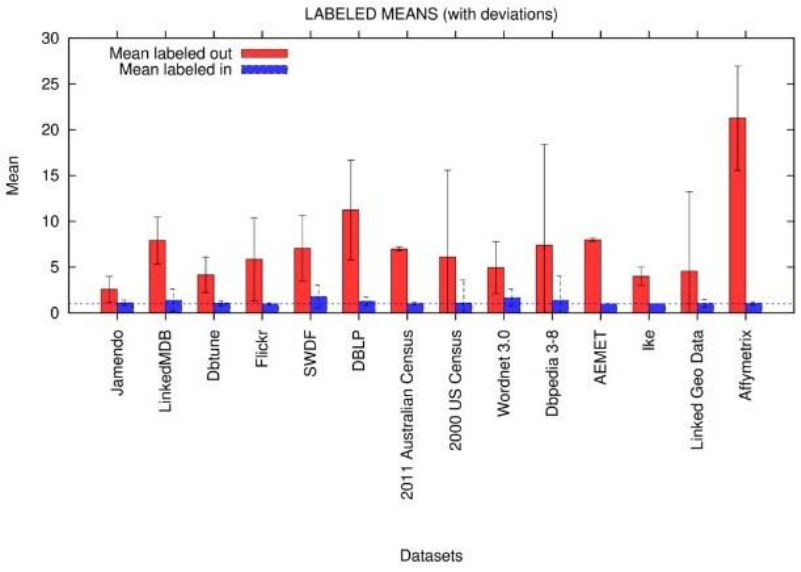

Figure 5 Mean labeled out- and in-degrees for the evaluation datasets.

The analysis of the mean labeled in-degree reveals an important conclusion. As shown in Figure 5, the average number of predicates related to a given object is very close to 1. This stands for specific "leave nodes" reached by only one different predicate.

Table 4 provides the maximum labeled out- and in-degrees for the analysed datasets and the ratio over the total number of predicates. The results for the maximum out- and in-degrees confirm the previous facts (the distribution is not skewed and the deviation is small), i.e., even in corner cases, few predicates are related to the same subject, and even less predicates are related to the same object.

The ratio of maximum degrees provided in Table 4 comes to similar conclusions. For instance, a ratio of *20%* in the labeled out-degree of *Wordnet* means that, at most, a subject is related to the *20%* of the predicates in the dataset. As expected, the smallest ratios correspond to the less structured datasets such as *Dbpedia* and *Linked Geo Data*.

Finally, it is worth noting that we studied the mean labeled degrees of the common subject-objects with respect to the values obtained without restrictions, obtaining similar results. The corner case was *Affymetrix,* which presents a significant reduction for subject-objects. One could argue that, in this case, general entities are detailed in depth whereas common subject-objects are simple nodes grouping discrete values and hence its smaller number of related predicates. Note that the mean labeled in-degree of common subject-objects remained close to *1*, i.e. the intermediate nodes (which are important for navigation) are also reached by a mean of one unique predicate, on average. Therefore, particular solutions could be designed in such case.

Table 4 *Values and ratios of the maximum labeled out- and in-degree for the experimental framework.*

	Dataset	Max. predicates per subjects		Max. predicates per object	
		Labeled out deg. $deg_L^-(G)$	*Ratio* $\frac{\lvert deg_L^-(G)\rvert}{\lvert P_G\rvert}$	*Labeled out deg.* $deg_L^+(G)$	*Ratio* $\frac{\lvert deg_L^+(G)\rvert}{\lvert P_G\rvert}$
Media	Jamendo	10	28.46%	5	19.23%
	LinkedMDB	31	13.96%	50	22.52%
	Dbtune	24	6.09%	93	23.60%
	Flickr Event Media	14	60.87%	5	21.74%
Publications	SWDF	21	15.91%	13	9.85%
	Faceted DBLP	18	66.67%	4	14.81%
Knowledge	Wordnet 3.0	17	20.00%	10	11.76%
	Dbpedia 3-8	480	0.83%	6,005	10.36%
Government	2011 Australian Census	7	26.92%	3	11.11%
	2000 US Census	104	24.24%	366	85.31%
Sensors	AEMET	12	52.17%	5	21.74%
	Ike	5	41.67%	1	8.33%
Geography	Linked Geo Data	76	0.42%	3,431	18.78%
Biology	Affymetrix	35	33.33%	5	4.76%

4.5. Partial and Direct Degrees

First of all, let us remember that partial out- and in-degrees reflect the presence of multivalued (*subject,predicate*) and (*predicate,object*) pairs respectively. Figure 6 shows the mean partial out- and in-degrees. As can be seen, the mean partial out-degree is slightly bigger than 1, revealing that the presence of multivalued (*subject,predicate*) pairs is not so frequent. In fact, the deviation is not pronounced (except for *Wordnet*) which denotes a uniform distribution. In contrast, the mean in-degree remains close to 1, but it presents bigger deviations. Almost all deviation extends the range of the figure and they have been erased for the sake of clarity. This fact denotes a pronounced skewed distribution of multivalued *(predicate,object)* pairs. That is, a large amount of different subjects are related to the same *(predicate,object)* (*e.g.* this can be the case of *rdf:type* and its related classes) while others pairs are related to few subjects, being 1 on average.

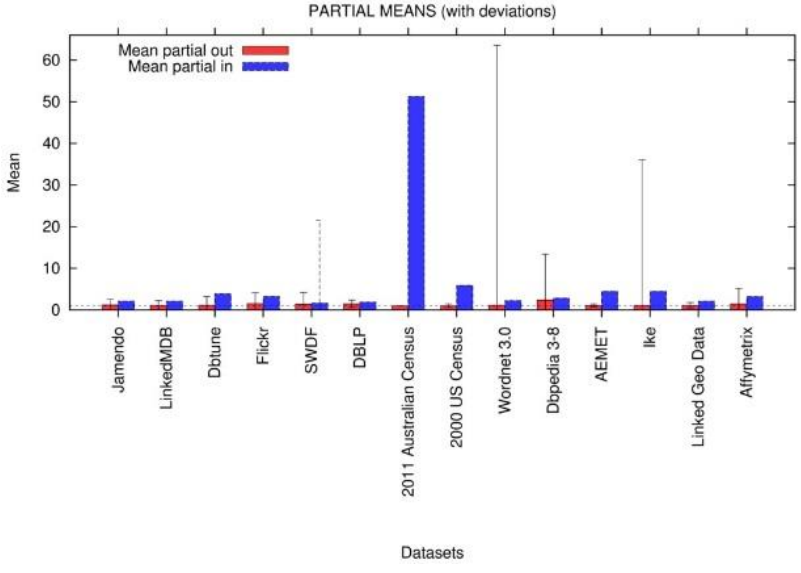

Figure 6 Mean partial out- and in-degrees for the evaluation datasets.

Next, we study the direct degrees, which measure the relationship between subjects and objects disregarding the presence of predicates. Figure 7 compares the mean out- and in-degrees and their respective mean direct degrees, showing that they have similar figures. That is, given a subject and an object, if they are related, only one predicate brings these nodes together, on average.

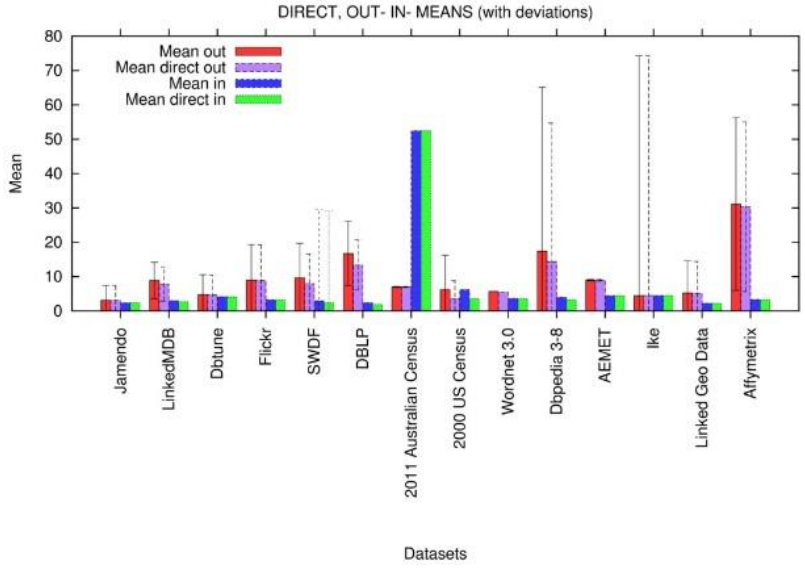

Figure 7 Mean direct degrees in comparison with mean out- and in-degrees for the evaluation datasets.

4.6. Predicate Degrees

In this section we study the predicate degrees, i.e the cardinality of predicates. We also detail their out- and in-degrees, that is, the objects and subjects related to each predicate. Figure 8 shows the mean predicate degrees for all datasets.

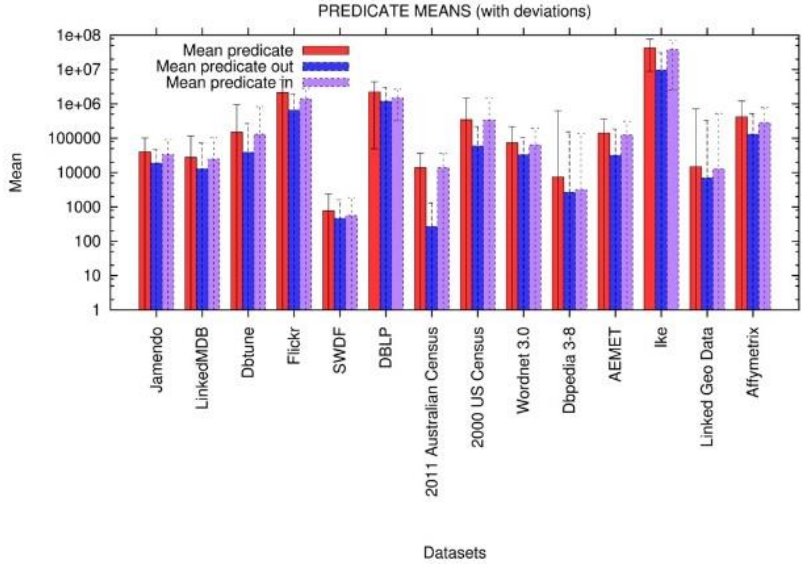

Figure 8 Mean predicate degrees for the evaluation datasets. The y-axis is in logarithmic scale.

First, note that predicate degrees are highly biased by the number of triples of each dataset. For instance, one could add more observations in *Ike* and the cardinality will be increased resp. In general terms, we can observe that the mean predicate out degree is slightly smaller than the corresponding mean in degree. That is, given a predicate at random, it is probably related with more subjects than objects. This fact is in line with previous labeled and partial measurements; subjects are more related to predicates than objects, and multivalued (*subject,predicate*) pairs are, when present, more infrequent than (*predicate, object*) pairs.

We then study, in the following, the distribution of predicates, as they can reveal different use patterns for the predicates. It is clear that no prior assumption can be made on predicate distribution as, in general terms, predicate distribution is tight to the information modelling. Nevertheless, in the studied datasets, we can roughly distinguish three types of patterns, illustrated in Figure 9 for the sensor, geography and biology domain:

- Clear power law distributions, where most predicates are present in a reduced number of triples, whereas few predicates are related to thousand or millions of triples. This corresponds to the definition of a power law distribution. We can find these very clear skewed distributions in cross-domain datasets such as *Dbpedia*, or datasets including information about a given domain but from diverse sources such as *Linked Geo Data* (Figure 9, bottom left). Due to the same reasons, these two datasets hold the higher numbers of predicates of all the evaluation datasets (see Table 2).
- Skewed distributions, i.e. some predicates are present rarely while others are frequently used, but not accurately fitting a power law distribution. This is the general tendency in most of the studied datasets, illustrated in *AEMET* (Figure 9, top left) and *Affymetrix* (Figure 9, bottom right).
- Flat distribution where mostly all predicates are present in every entity. In such case, the predicates are in the same region as they participate in a similar range of triples, exemplified in (Figure 9, top right).

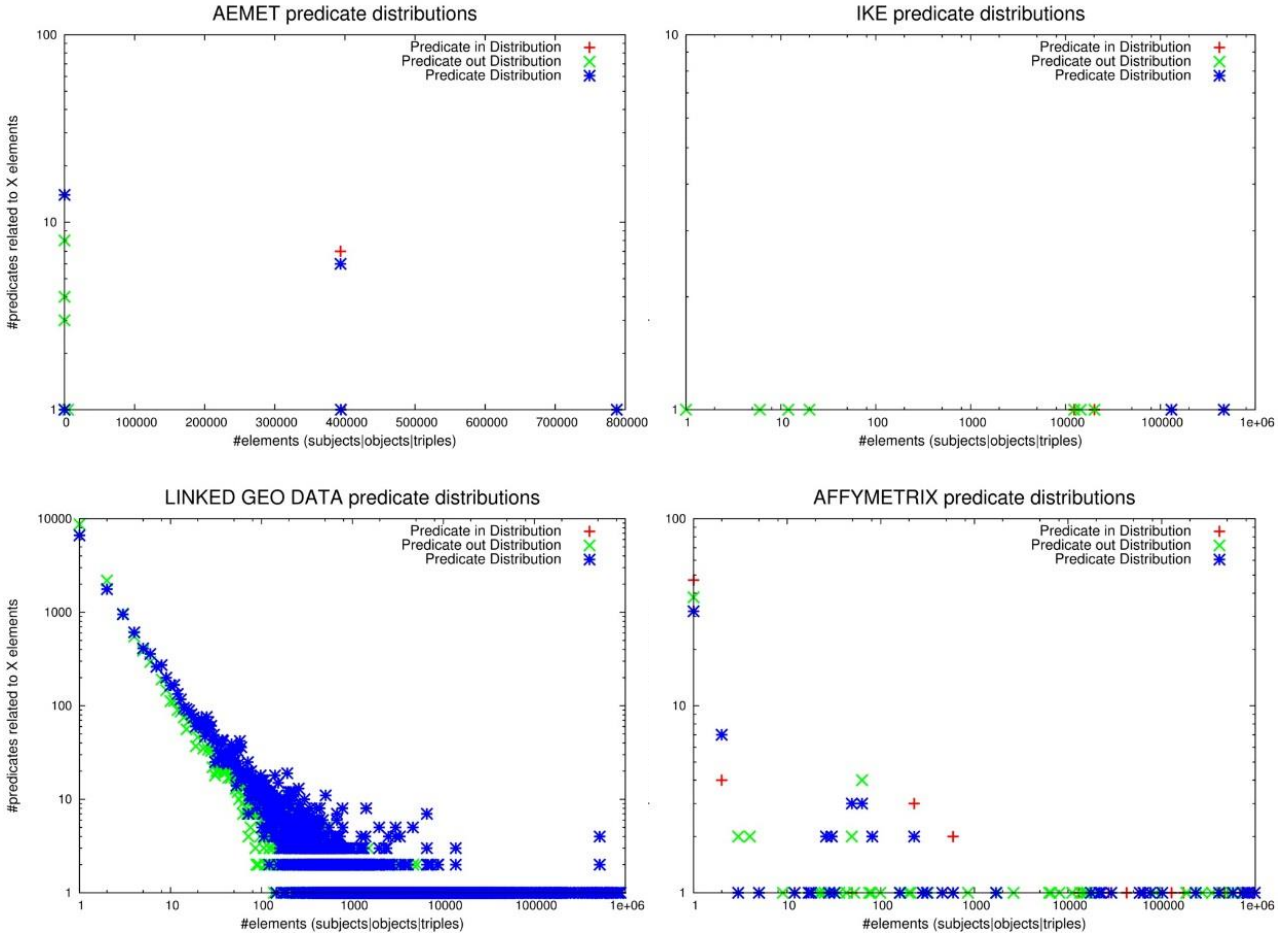

Figure 9 Predicate degree distribution (sensors, geography and biology), in logarithmic scale.

4.7. Study of Predicate Lists

Table 5 presents the number and repetition ratio of predicate lists, for all subjects as well as restricted to typed subjects. As can be seen, in both cases the number of different predicate lists is spectacularly lo. For instance, in *Jamendo*, which holds *26* predicates, only *43* different lists are present (*999.872‰* of the lists are repetitions). This fact remains true event in those cross-domain datasets with more predicates and thus different entities such as *Dbpedia*.

Table 5 *Number and ratio of predicate lists for all subjects (left) and restricted to typed subjects (right).*

	Dataset	All subjects		Typed subjects	
		Predicate lists $\lvert L(G)\rvert$	*Repetition Ratio* $(1-\frac{\lvert L(G)\rvert}{\lvert S_G\rvert})$	*Predicate lists* $\lvert L(G)\rvert$	*Repetition Ratio* $(1-\frac{\lvert L(G)\rvert}{\lvert S_G\rvert})$
Media	Jamendo	43	999.872‰	41	999.859‰
	LinkedMDB	8,459	987.818‰	8,442	987.314‰
	Dbtune	963	999.922‰	782	999.922‰
	Flickr Event Media	25	999.996‰	22	999.987‰
Publications	SWDF	364	965.254‰	341	961.754‰
	Faceted DBLP	254	999.929‰	254	999.929‰
Knowledge	Wordnet 3.0	872	999.208‰	868	999.007‰
	Dbpedia 3-8	1,309,392	947.184‰	1,152,617	712.413‰
Government	2011 Australian Census	14	999.730‰	14	999.730‰
	2000 US Census	106	999.996‰	-	-
Sensors	AEMET	5	999.987‰	5	999.987‰
	Ike	5	1,000.000‰	4	1,000.000‰
Geography	Linked Geo Data	220,902	995.745‰	219,015	995.562‰
Biology	Affymetrix	9,434	993.365‰	9,424	993.369‰

Next, we study the number of repetitions per list, and their distribution. This mean has been defined as the mean predicate list degree (Definition 18), and the results are shown in Figure 10.

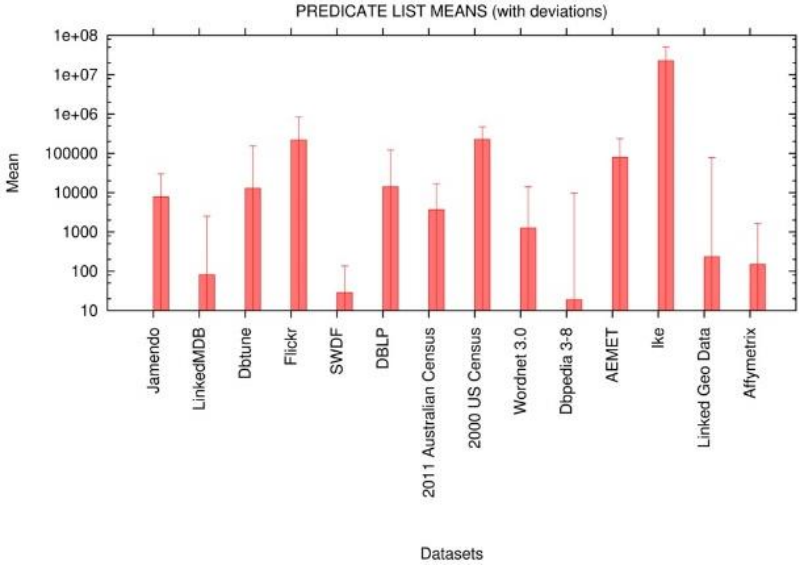

Figure 10 Mean predicate list degree for the evaluation datasets.

The results of the number of repetitions per list are in line with the presented repetition ratio. Nevertheless, it is important to note that, as for the predicate cardinality, these results can be highly biased by the number of triples (the y-axis is in logarithmic scale). In plain words, the more triples are present in the dataset, the higher can be the number of repetitions. Intuitively, one could expect that these distributions would correspond to the predicate distributions presented in the previous Section 4.6. That is, if a skewed distribution is present in predicates, the same result could be found in predicate lists. In turn, if all predicates participate in similar number of triples (uniform distribution), the same shape is present in predicate lists. Our evaluation analysed the predicate list distribution and showed that both assumptions remain true for the studied datasets, exemplified in Figure 11 for *Linked Geo Data* and *IKE*.

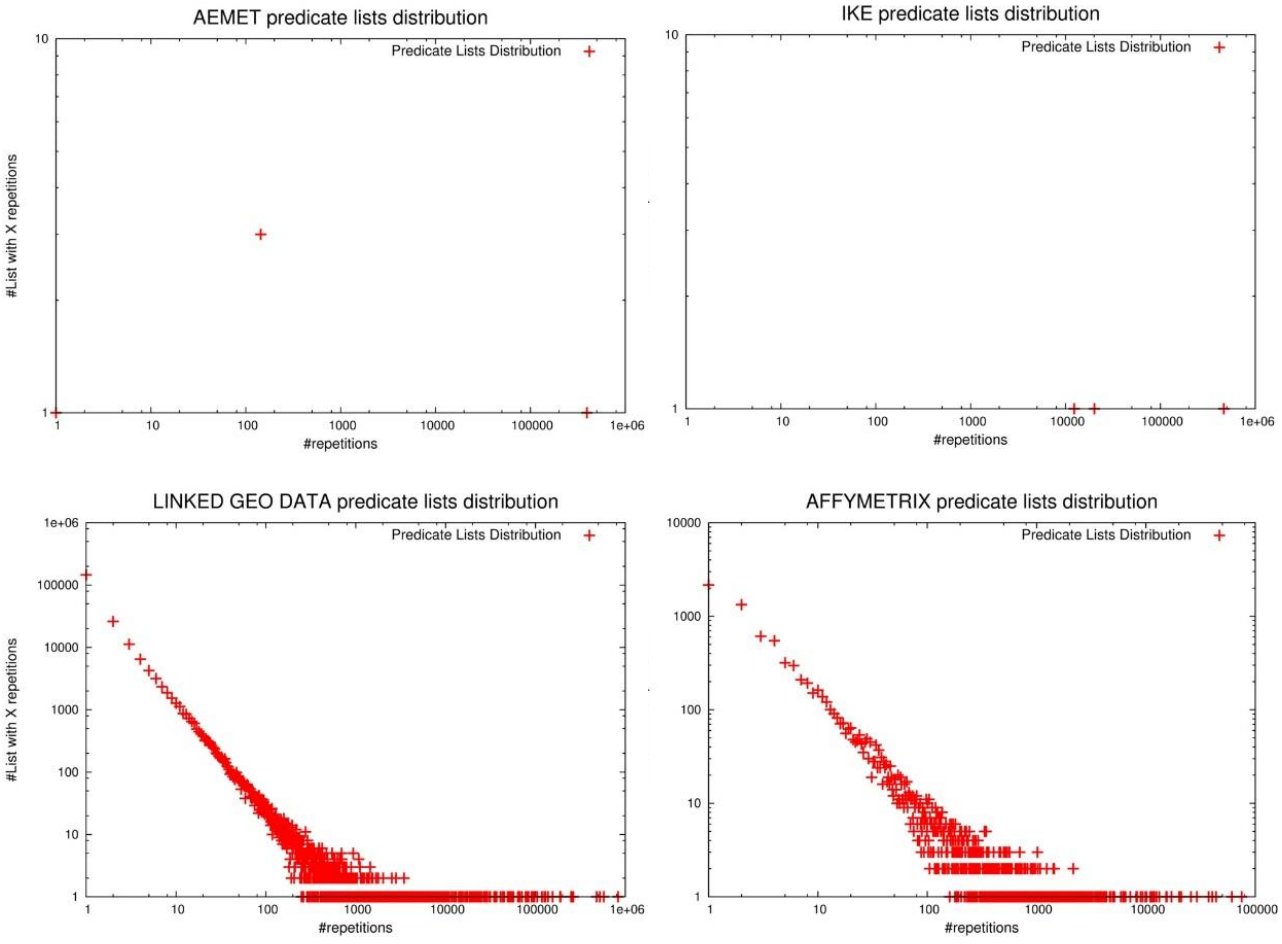

Figure 11 Predicate list degree distribution (sensors, geography and biology), in logarithmic scale.

Finally, we study the number of different lists per predicate, on average. This is shown in Figure 12 (in logarithmic scale) which shows a significant low number of different lists in which a predicate is present. Note that if a predicate was related to one or two lists, given a predicate it is almost direct to know its peer predicates for any subject or object, even for the biggest datasets. In other words, the nearer this means is to 1, the easier could be to discern the concrete list given a predicate, even in the biggest datasets. The highest figures are obviously obtained for those datasets with more predicate lists, but they remain proportionally small to the number of lists.

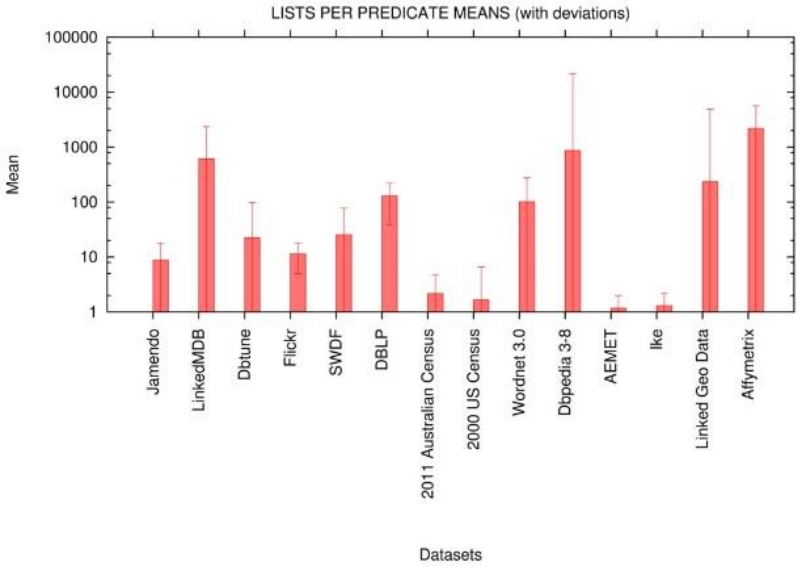

Figure 12 Mean list per predicate degree for the evaluation datasets, in logarithmic scale.

4.8. Study of Classes and Typed Subjects

We finish our evaluation with a brief study on typed entities given the importance for other uses such as reasoning (see the potential uses of these metrics in Section 3.6). Table 6 shows the resulting number of classes, typed subjects and the ratio of these typed subjects over the total subjects. First, one could expect that the larger is the dataset, the more classes are included. However, it is worth remembering that RDF holds a relaxed schema, hence this assumption can result completely false. In other words, a "small" dataset such as *Jamendo* or SWDF can include more classes than the bigger *Flickr* or *AEMET*. Thus, the number of classes and typed subjects is completely biased by the data modelling and the domain/s involved in the dataset.

Table 6 *Number of classes, typed subjects and its ratio for the experimental framework.*

	Dataset	Classes	Typed Subjects ($\lvert S_G^C \rvert$)	Ratio ($\frac{\lvert S_G^C \rvert}{\lvert S_G \rvert}$)
Media	Jamendo	11	290,291	86.42%
	LinkedMDB	53	665,441	95.83%
	Dbtune	64	10,042,747	80.96%
	Flickr Event Media	3	1,690,338	30.79%
Publications	SWDF	62	8,916	85.11%
	Faceted DBLP	14	3,591,091	100.00%
Knowledge	Wordnet 3.0	25	873,986	79.42%
	Dbpedia 3-8	351	4,007,892	16.17%
Government	2011 Australian Census	15	51,768	100.00%
	2000 US Census	0	0	0.00%
Sensors	AEMET	5	394,289	100.00%
	Ike	12	114,471,666	99.99%
Geography	Linked Geo Data	1,081	49,352,200	95.06%
Biology	Affymetrix	3	1,421,291	99.97%

With this assumption in mind, we can find in the results that the number of classes remain proportionally small with respect of the number of triples and entities. This is an obvious result as classes model common semantic types of entities, and this distinction should be limited. However, the ratio of typed subjects draws more interesting results.

Table 6 reflects three types of modelling:

- Non-typing: in this case no types are used, such as the *2000 US Census*.
- Small-medium typing: datasets in which types are used around one of every four subjects (≅ 25%). In our study, we find two cases, *Flickr* (31%) and *Dbpedia* (16%), matching this scenario.
- Extensive-typing: most subjects are typed. This is the case of most datasets in our study, ranging from 79% to 100% of typed subjects.

Next, we extend our previous study on predicates, performing a mean of predicate lists per class (see Definition 21). This is represented in Figure 13. Note that the mean is exactly 1 (with no deviation) for *AEMET* and *Ike*.

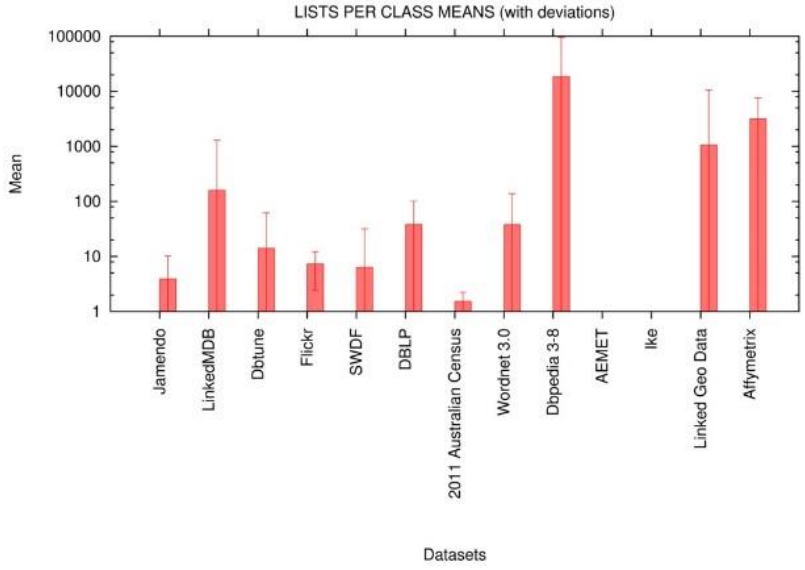

Figure 13 Mean lists per class for the evaluation datasets. The y-axis is in logarithmic scale.

The mean figures show that, seven of thirteen datasets hold a mean of less than ten predicate lists per class, and it remains valid independently of the size of the dataset. This means that, given a class, we can automatically state that all subjects of this class are described with one of ten variations of predicates, on average. Another three datasets range between 10 and 100 lists per class (which remains still small). The three datasets with more different lists, obviously present more lists per class (up to 19,000 for *Dbpedia*). Nevertheless, in these latter cases the deviation is also high, hence we can also find classes with much lesser variants.

Finally, we present in Figure 14 a brief comparison of mean out-degrees for typed subjects with respect to all subjects. We restrict to those datasets having small-medium typing (as defined above), *Flickr* and *Wordnet*, as in the extensive-typing case both means are similar and the comparison makes no sense. We extend the range of the y-axis to show that both means and deviations are comparable. Nevertheless, typed subjects are actually described with slightly more triples than those subjects without restrictions, on average. This can be seen as a way of providing detailed descriptions for these special nodes, of particular interest to navigate and organize the graph.

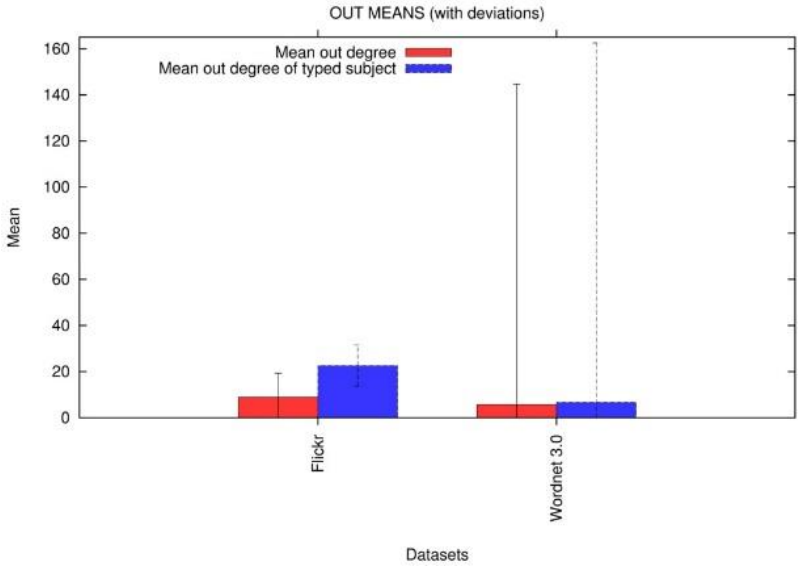

Figure 14 Mean out degree for the evaluation datasets in comparison with typed subjects.

5. Discussion

5.1. Contributions

In this work we have studied and characterized the real structure of RDF datasets. We have motivated our purpose in the sparingly number of previous empirical studies and the few parameters considered. Then, we propose and define novel metrics for RDF aimed at characterizing real-world RDF data. Our initial purpose was to provide a toolkit of parameters that could both (i) determine common features in most RDF datasets and (ii) become a useful handbook when developing or optimizing RDF data structures, indexes and other related technologies.

The proposed metrics cover a wide spectrum of parameters. First, the RDF dataset is regarded as a graph labeled with predicates, and we give metrics to characterize the subject (out-) and object (in-) distributions. We measure their degree (out- and in-degrees respectively), the presence of multivalued pairs (partial degree), the number of different predicates per node (labeled degree) and the direct relationships between nodes disregarding labels (direct degree).

Then, we characterize the distribution of predicates, which is of great importance as they hold the semantics of the datasets. We define their cardinality (predicate degree), and the distribution of subjects and objects per predicate (predicate in and out-degree). This later is equivalent to describe the domain and range of each predicate.

We consider the repetitions of nodes playing different roles, hence common ratios are defined: subject-object, subject-predicate and predicate-object. Given the importance of the first ones as hubs in the navigation of the graph, we propose to characterize the subject and object degrees restricted to common subject-objects.

Agreeing that the list of predicates per subject can be repeated in several subjects, we then focus on parameterize these list and their repetitions. We define a ratio of repeated predicate lists, the cardinality of each list (predicate list degree) and the number of lists in which each predicate takes part (lists per predicate degree).

Finally, we make a special distinction of typed subjects, as they could share commonalities. We count the number of classes, typed subjects and their ratio over the total number of subjects. We also define the number of different predicate lists per class (lists per class degree) and consider the subject and predicate list degrees restricted to typed subjects.

5.2. Result Summary and Applications

We established an evaluation framework consisting in fourteen datasets trying to cover a wide range of different datasets. The following summary of conclusions can be drawn from the evaluation results:

- All datasets show a skewed structure on subjects and even more remarkable on objects, with very clear power-law distributions. This is, of course, the perfect scenario to allow RDF compressors to obtain high compression rates. In turn, RDF stores can also leverage these statistical properties to adapt their query plans to the selectivity of the involved entities. Note that, in general, the mean degree of subjects and objects (established in 10 triples in the studied datasets) is not representative as these skewed distributions result in a high standard deviation.
- As we expected, subject-predicate and predicate-object ratios are almost negligible and the number of classes remains proportionally small w.r.t the number of triples and entities in the dataset. This result highlights that RDF stores can indeed make use of dedicated structures to manage schema and data instances in order to optimize the space of the representation and to speed up specific schema retrieval operations and reasoning processes.
- Complementing the previous finding, our evaluation also revealed that most datasets are extensively typed (more than 80% of the subjects have an *rdf:type*), and that each class tends to appear in few predicate lists (less than 10 predicate lists per class), independently of the dataset size. This states that RDF compressing and indexing techniques can treat *rdf:type* as a frequent predicate and optimize its representation (e.g. assigning less bits to encode it or even representing it implicitly). Then, it implies that the type of the subject univocally determines the predicates to which the subject is related, so that applications (e.g. serializations, RDF stores or visualizers/browsers) can use this association for multiple purposes, such as optimizing the underlying structures, curating datasets by suggesting changes, or inferring implicit information.
- Subject-object was identified as the most frequent path constructor, which can constitute more than 60% of the entities. This result directly affects how RDF compressors and RDF stores encode these intermediate nodes. In general, a dictionary encoding is used to assign one unique integer ID to each subject, predicate and object in the graph, hence minimizing the space overheads of long repeated strings and managing a much more efficient graph of IDs. Thus, a high ratio of subject-objects points out that these dictionary-based techniques could assign a unique ID to each of these nodes, with a two-fold objective. First, strings are encoding just once (instead of represented twice with the subject and object roles), which results in significant space savings. Then, an efficient identification and filtering of these particular subject-object IDs can speed up the navigation of the graph. Nonetheless, results also show that the design of the dataset has a strong influence in the presence of such intermediate nodes.
- Most datasets show that each subject is described with less than 10 different predicates, on average, remaining true if we restrict to common subject-objects. In addition, the number of different predicate lists is spectacularly low. These results suggest again that RDF managing systems can make use of structural clusters for different purposes. For instance, RDF compressors and RDF stores can assign one ID to each different predicate list and can efficiently encode the predicates related to a given subject by referring to the ID of the predicate list. Furthermore, the distribution of predicate lists is also skewed, hence its representation is highly compressible. Our analysis also reveals that each predicate participates in a small number of predicate lists. Thus, retrieval mechanisms can index such relations to boost performance when retrieving all the triples for a given predicate, filtering in which lists the given predicate may appear.
- Surprisingly, our evaluation highlights that the number of predicates related to a given object is very close to 1. This suggests that systems can isolate the use of each predicate and establish a range of application (object values for each predicate) with none or few intersections. This isolation can improve the codification of the graph and its space needs (a smaller range implies less bits to encode each value).
- Regarding multivalued pairs, experiments show that each (*subject*,*predicate*) is mostly related to a unique object, and each (*predicate*,*object*) is often related to one subject, but the high standard deviation of this latter denotes a pronounced skewed distribution. These results suggest that the adjacency list encoding of the first one requires less bits that the latter, as one can use an encoding considering that each (*subject*,*predicate*) pair is related to one object by default, and only use a special mark to know those few pairs that are related to more objects.
- Finally, the results for predicate degrees state that, on average, each predicate is related with more subjects than objects. This impacts query resolution plans, where a query with a bounded subject, (S,P,?o), should be promoted over a bounded object, (?s,P,O), as the first one produces less intermediate results than the latter. In contrast, an RDF index by predicate and object (referred to as POS) may require less space than one by predicate and subject (referred to as PSO), as the number of objects per predicate is smaller than the subjects per predicate.

5.3. Future work

We expect that these metrics and observations can provide insights to take advantage of some of the revealed features to develop and optimize better dataset designs, visualizations, efficient RDF data structures, indexes (in particular, structural indexes) and compression techniques. The concrete optimizations and decisions are subject of each particular scenario and are out of the scope of this study, nonetheless we have introduced potential application scenarios.

In particular, as the number of predicates per object is close to 1, this stands for a specific treatment of these "leave nodes" for each predicate. Thus, approaches such as a specific compression over vertical partitioning can obtain important results. In turn, the number of few predicates per subject and their distribution (labeled out degree) is a clear indicator of the presence of star-shaped nodes. Together with the characterization of intermediate nodes, this could serve query suggestion and visualization purposes. In particular, it is highly remarkable that intermediate nodes are reached by a mean of one predicate, reducing the number of predicates which connects different parts of the graph.

The family of indexing techniques following vertical partitioning can consider also the predicate distributions and, potentially, make use of these metrics to optimize the resolution of complex queries.

Predicate lists and the characterization of classes would serve several purposes such as visualization, structural indexing for querying and reasoning. We would like to remark the massive repetition of predicate lists, in general, and the low number of predicates per class in particular. This may help in determining structural indexes for such purposes.

Finally, as a complementary effort, we expect that future studies on the state of the art of the linked open data cloud will consider our fine-grained metrics to help categorize different design patterns.

Notes

1. http://stats.lod2.eu/
2. http://linkeddata.org/
3. http://xmlns.com/foaf/spec/
4. http://dbpedia.org/
5. http://bio2rdf.org/
6. http://any23.apache.org/
7. ftp://nassdataweb.infor.uva.es/RDFmetrics/datasets/

Funding

The work presented in this paper has been part-funded by Austrian Science Fund (FWF): M1720-G11 and Ministerio de Economía y Competitividad, Spain: TIN2013-46238-C4-3-R. C. Gutiérrez was partly funded by Millennium Nucleus Center for Semantic Web Research under Grant NC120004. Authors are also grateful to the KEYSTONE COST Action IC1302.

References

[1] Bizer C, Heath T, and Berners-Lee T. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts 2009, pp. 205-227.

[2] Manola F, Miller E, and McBride, B. RDF 1.1 primer. W3C Working Group Note 24 June 2014. https://www.w3.org/TR/rdf11-primer/.

[3] Harth A, Umbrich J, Hogan A, Decker S. YARS2: A Federated Repository for Querying Graph Structured Data from the Web. In: Proceedings of the International Semantic Web Conference 2007, pp. 211–224.

[4] Neumann T, and Weikum, G. The RDF-3X engine for scalable management of RDF data. The VLDB Journal 2010, vol. 19, pp. 91–113.

[5] Urbani J, Maassen J, Drost N, Seinstra F, and Bal H. Scalable RDF data compression with MapReduce. Concurrency and Computation: Practice and Experience 2013, 25(1), pp. 24-39.

[6] Fernández J D, Martínez-Prieto M A, Gutiérrez C, Polleres A, and Arias M. Binary RDF representation for publication and exchange (HDT). Web Semantics: Science, Services and Agents on the World Wide Web 2013, vol. 19, pp. 22-41.

[7] Potter A, Motik B, and Horrocks, I. Querying Distributed RDF Graphs: The Effects of Partitioning. In 10th International Workshop on Scalable Semantic Web Knowledge Base Systems 2014, p. 29.

[8] Cai M, and Frank, M. RDFPeers: a scalable distributed RDF repository based on a structured peer-to-peer network. In Proceedings of the 13th international conference on World Wide Web 2004, pp. 650-657.

[9] Ding, L, and Finin, T. Characterizing the Semantic Web on the Web. In Proceedings the International Semantic Web Conference 2006, pp. 242–257.

[10] Theoharis Y, Tzitzikas Y, Kotzinos D, and Christophides V. On Graph Features of Semantic Web Schemas. IEEE Transactions on Knowledge and Data Engineering 2008, 20(5), 692–702.

[11] Bachlechner, D, and Strang T. Is the Semantic Web a Small World? In Proceedings of the International Conference on Internet Technologies and Applications 2007, pp. 413–422.
[12] Ge W, Chen J, Hu W, and Qu Y. Object Link Structure in the Semantic Web. In Proceedings of the Extended Semantic Web Conference 2010, pp. 257–271.
[13] Gil Y, and Groth P. LinkedDataLens: linked data as a network of networks. In Proceedings of the International Conference on Knowledge Capture 2011, pp. 191–192.
[14] Brickley D, and Guha R V. RDF Schema 1.1. W3C Recommendation, 25 February 2014. https://www.w3.org/TR/rdf-schema/.
[15] W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012. https://www.w3.org/TR/owl2-overview/.
[16] Erdős P, and Rényi A. On random graphs. Publicationes Mathematicae 6, 1959, pp. 290–297.
[17] Albert R, Jeong H, and Barabasi, A L. Diameter of the World Wide Web. Nature 1999, vol. 401(February), pp. 130–131.
[18] Redner S. How popular is your paper? An empirical study of the citation distribution. European Physical Journal B 1998, vol. 4(2), pp. 131–134.
[19] Jeong H, Mason S P, Barabasi, A L, and Oltvai, Z N. Lethality and centrality in protein networks. Nature 2001, vol. 411(6833), pp. 41–42.
[20] Zhang H. The scale-free nature of semantic web ontology. In Proceedings of the World Wide Web Conference 2008, pp. 1047–1048.
[21] Hu W, Chen J, Zhang H, and Qu Y. How Matchable Are Four Thousand Ontologies on the Semantic Web. In Proceedings of the Extended Semantic Web Conference 2011. pp. 290–304.
[22] Watts D J, Networks, Dynamics, and the Small World Phenomenon. American Journal of Sociology 1999,105(2), pp. 493–527.
[23] Milgram S. The small world problem. Psychology Today 1967, vol. 2(1), pp. 60–67.
[24] Gil R, and García R. Measuring the Semantic Web. AIS SIGSEMIS Bulletin 2004, vol. 1(2), pp. 69–72.
[25] Adamic L A. The Small World Web. In Proceedings of the European Conference on Digital Libraries 1999, pp. 443–452.
[26] Guns R. Unevenness in Network Properties on the Social Semantic Web. Scalable Computing: Practice and Experience 2008, vol. 9(4), pp. 271–279.
[27] Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata, R. et al. Graph structure in the Web. Computer Networks 2000, vol. 33(1-6), pp. 309–320.
[28] Cheng G, Ge W, and Qu Y. Falcons: searching and browsing entities on the semantic web. In Proceedings of the World Wide Web Conference 2008, pp. 1101–1102.
[29] Hogan A, Polleres A, Umbrich J, and Zimmermann A. Some entities are more equal than others: statistical methods to consolidate Linked Data. In Workshop on New Forms of Reasoning for the Semantic Web: Scalable & Dynamic 2010.
[30] Hogan A, Zimmermann A, Umbrich J, Polleres A, and Decker S. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. Web Semantics: Science, Services and Agents on the World Wide Web, 2012, vol. 10, pp. 76–110.
[31] Hogan A, Harth A, Passant A, Decker S, and Polleres A. Weaving the Pedantic Web. In Linked Data on the Web 2010.
[32] Mihindukulasooriya N, Poveda-Villalón M, García-Castro R, Gómez-Pérez A. Loupe - An Online Tool for Inspecting Datasets in the Linked Data Cloud. In Proceedings of International Semantic Web Conference 2015, CEUR 1486, paper 113.
[33] Perez J, Arenas M, and Gutiérrez C. Semantics and Complexity of SPARQL. ACM Transactions on Database Systems 2009, vol. 34(3), pp. 1-45.
[34] Gutiérrez C, Hurtado C, Mendelzon A O, and Perez J. Foundations of Semantic Web Databases. Journal of Computer and System Sciences 2011, vol. 77, pp. 520–541.
[35] Hayes J, and Gutiérrez C. Bipartite Graphs as Intermediate Model for RDF. In Proceedings of the International Semantic Web Conference 2004, pp. 47–61.
[36] Salomon D. Data Compression: The Complete Reference. Springer-Verlag London Limited. 2007.
[37] Abadi D, Adam M, Madden S R, and Hollenbach K. Scalable Semantic Web Data Management Using Vertical Partitioning. In Proceedings of the Very Large Data Bases Conference 2007, pp. 411–422.
[38] Atre M, Chaoji V, Zaki M J, and Hendler J A. Matrix "Bit" loaded: a scalable lightweight join query processor for RDF data. In Proceedings of the World Wide Web Conference 2010, pp. 41–50.
[39] Khatchadourian S, and Consens M P. ExpLOD: Summary-Based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In Proceedings of the Extended Semantic Web Conference 2010, pp. 272–287.
[40] Campinas S, Perry T E, Ceccarelli D, Delbru R, and Tummarello G. Introducing RDF Graph Summary with Application to Assisted SPARQL Formulation. In Proceedings of the 23rd International Workshop on Database and Expert Systems Applications 2012, pp. 261-266.
[41] Tran T, Ladwig G, and Rudolph S. Managing Structured and Semistructured RDF Data Using Structure Indexes. IEEE Transactions on Knowledge and Data Engineering 2013, vol. 25(9), pp. 2076–2089.
[42] Hernández-Illera A, Martínez-Prieto M A, and Fernández J D. Serializing RDF in compressed space. In Proceedings of Data Compression Conference 2015, pp. 363-372.
[43] Carothers G, Seaborne A. RDF 1.1 N-Triples. W3C Recommendation 25 February 2014. https://www.w3.org/TR/n-triples/.