Semantic Web 0 (0) 1
IOS Press

Charaterizing RDF Graphs through Graph-based Measures - Framework and Assessment

Matthäus Zloch [a,c,*], Maribel Acosta [b], Daniel Hienert [a], Stefan Conrad [c] and Stefan Dietze [a,c]
[a] *GESIS - Leibniz-Institute for the Social Sciences, Cologne, Germany*
E-mails: matthaeus.zloch@gesis.org, daniel.hienert@gesis.org, stefan.dietze@gesis.org
[b] *Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany*
E-mail: maribel.acosta@kit.edu
[c] *Institute DBS, Heinrich-Heine University, Düsseldorf, Germany*
E-mails: matthaeus.zloch@hhu.de, stefan.conrad@uni-duesseldorf.de, stefan.dietze@hhu.de

Abstract. The topological structure of RDF graphs inherently differs from other types of graphs, like social graphs, due to the pervasive existence of hierarchical relations (TBox), which complement transversal relations (ABox). Graph measures capture such particularities through descriptive statistics. Besides the classical set of measures established in the field of network analysis, such as size and volume of the graph or the type of degree distribution of its vertices, there has been some effort to define measures that capture some of the aforementioned particularities RDF graphs adhere to. However, some of them are redundant, computationally expensive, and not meaningful enough to describe RDF graphs. In particular, it is not clear which of them are efficient metrics to capture specific distinguishing characteristics of datasets in different knowledge domains (e.g., *Cross Domain* vs. *Linguistics*). In this work, we address the problem of identifying a minimal set of measures that is efficient, essential (non-redundant), and meaningful. Based on 54 measures and a sample of 280 graphs of nine knowledge domains from the Linked Open Data Cloud, we identify an essential set of thirteen measures, having the capacity to describe graphs concisely. These measures have the capacity to present the topological structures and differences of datasets in established knowledge domains.

Keywords: RDF Graph, Graph Topology, Graph Measures, Measure Assessment, RDF Graph Profiling

1. Introduction

Characteristics of RDF graphs can be captured through descriptive statistics using graph-based measures.

Understanding the topology of RDF graphs can guide and inform the development of, e.g., synthetic dataset generators, sampling methods, profiling tools, dataset discovery, index structures, or query optimizers. Solutions in the aforementioned research areas rely on *effective* measures and statistics, in order to be compliant with real-world situations and to return appropriate results.

RDF graphs have a distinct topology from other graphs, like social graphs or computer networks, due to the pervasive existence of hierarchical relations: relations within the ABox (assertional statements - the data) are complemented by relations within the TBox (terminological statements - schema definitions, e.g., rdfs:subClassOf) as well as between ABox and TBox. rdf:type is probably the most famous example adhering to almost every description of a resource in an RDF dataset. These particularities are directly reflected in one RDF graph's topology and lead to, e.g., higher overall connectivity and existence of redundant structural patterns in the graphs, and as such, they cannot be captured with ordinary measures. In addition to known measures from the field of

*Corresponding author. E-mail: matthaeus.zloch@gesis.org.

network analysis [1, 2], such as the number of vertices/ edges and the distribution of vertex degrees, there has been some effort to define measures to characterize RDF graphs [3], in order to capture the aforementioned particularities RDF graphs involve.

Problem Statement

Computing arbitrary graph measures for RDF graphs is computationally expensive. Measures like diameter (the longest shortest path in a graph), clustering coefficient (tendency of the graph to build clusters), or the mean repetitive distinct predicate set usage per subject, e.g., involve a degree of complexity and are costly in terms of computation time (depending of the size of the graph, i.e., number of vertices/edges). Focusing on an efficient set of descriptive measures helps RDF profiling tools to speed up the process and to create *concise* descriptions of RDF graphs.

The main objective of this paper is to identify such an essential set of measures. We aim to identify a set of meaningful, efficient, and non-redundant measures, for the goal of describing RDF graph topologies more accurately and facilitating the development of the aforementioned solutions. An *efficient* measure is considered to be discrete and adding extra value in describing a graph, without being dependent on another measure. Its existence contributes to the conciseness of a graph's description.

Approach and Methodology

In order to gain an understanding of measure effectiveness and identify optimal graph measures, we investigate 54 distinct graph measures and apply feature engineering techniques on various tasks. Our study bases on 280 RDF datasets sampled from all categories of the Linked Open Data Cloud[1] (LOD Cloud) late 2017, and values of about 54 (RDF) graph-based measures.

We follow a three-stage approach. First, we investigate feature redundancy by computing feature correlations among all measures and apply feature selection methods, to eliminate redundant and non-effective measures. For the resulting set of non-redundant measures, we study measure variability in terms of statistical tests across and within categories, i.e., the nine distinct knowledge domains provided by the LOD Cloud. Finally, we assess measure performance concerning a measure's capacity to discriminate dataset categories in binary classification tasks, using state-of-the-art machine learning models.

The experiment results show that a large proportion of graph measures are redundant, in terms of that they do not add value to describe RDF graphs. Only the essential set of *thirteen* measures have the capacity to describe RDF graphs concisely. Moreover, characteristics of RDF graphs vary notably across knowledge domains, which is well reflected in the evaluation of measure impact when it comes to discriminating RDF graphs by knowledge domain.

Contributions and Structure

This work is considered an extension of a recently published paper [2][2].

Whereas key contributions of [2] include (a) a framework for efficiently computing graph measures and (b) an initial application of such measures to datasets of the LOD cloud, this work is an extension through the following contributions:

- Formal definitions of 28 graph measures in terms of RDF graphs (**§ 3**),
- Implementation of 31 RDF graph measures formally defined in [3], as an extension of the software framework[3], and
- an update of the website as a browsable version[4] for all datasets that were analyzed, with values from the measure computation.
- A graph-based analysis of a mixed set of fifty-four graph and RDF graph measures, obtained from a sample of 280 datasets from the LOD Cloud (**§ 4**).
- Identification of an efficient set of measures through feature engineering techniques, in order to retrieve concise descriptions about RDF graphs (**§ 5.1**).
- A report about topological differences of real-world RDF datasets within distinct categories (**§ 5.2**).
- An analysis of (RDF) graph measure performance, concerning their capacity to discriminate dataset categories (**§ 5.3**).

[1] https://lod-cloud.net/

[2] In order for this paper to be self-contained, please note that we have re-used some paragraphs, especially for the related work in Section 2, the textual descriptions of graph measures in Section 3.2, and for the description about the acquisition of RDF datasets from the LOD Cloud in Section 4.2.1.

[3] https://doi.org/10.5281/zenodo.2109469

[4] https://data.gesis.org/lodcc/2017-08/

- Based on our observations, we identify relevant measures or graph invariants that characterize graphs in the Semantic Web.

2. Related Work

The RDF data model imposes unique characteristics that are not present in other graph-based data models. Therefore, we distinguish between works that analyze the structure of RDF datasets in terms of RDF-specific measures and measures of graph invariants.

Many of the research related can be considered profiling approaches. An *RDF dataset profile* or *RDF summary graph* is a quantitative representation of an RDF dataset in terms of its features (characteristics) adhering at instance- and schema-level [4]. Profiling in this context means the activity of extracting such features from RDF datasets. Thus, some of the works mentioned appear in research activities in this domain of research [4, 5]. Creating an RDF summary graph aims at building concise overviews of the data in RDF knowledge bases [5], in order to optimize, for example, querying and processing times for SPARQL endpoints [6, 7], rather than aiming at extracting information about its topology.

RDF-specific Analyses

This category includes studies about the general structure and quality of RDF graphs at instance-, schema-, and metadata-levels. Schmachtenberg et al. [8] present the status of RDF datasets in the LOD Cloud in terms of size, linking, vocabulary usage, and metadata. LODStats [9] and the large-scale approach DistLODStats [10] report on descriptive statistics about RDF datasets on the web, including the number of triples, RDF terms, properties per entity, and usage of vocabularies across datasets. ExpLOD [11] generates summaries and aggregated statistics about the structure of RDF graphs, e.g., sets of used properties or the number of instances per class. In addition, [12] presents an approach for extracting structured topic profiles of RDF datasets from dataset samples. ProLOD++ [13, 14] is an online tool which profiles any RDF dataset. It reports on, for example, frequencies and distributions of subjects, predicates, objects, ratio of incoming/outgoing links, and performs pattern analysis on object values. It enables "to perform further analysis only on subsets of the dataset that correspond to clusters"[14]. Loupe [15], a "comprehensive linked data profiling tool", provides a RESTful web service for profiling SPARQL engines. The API reports on vocabulary, class, and property usage and cardinalities, and facilitates the analysis of implicit data patterns. Hogan et al. [16] study the distribution of RDF terms, classes, instances, and datatypes to measure the quality of public RDF data.

The quality aspect of Linked Open Data has been subject to some recent studies. Debattista et al. assessed the quality of metadata and dataset availability, investigating datasets from the LOD Cloud 2014 [17] and early 2019 [18]. Haller et al. [19] investigated different types of links, i.e., contained in the ABox and TBox, exposed by 430 datasets in the LOD Cloud.

A recent study provides a comprehensive overview of "available methods and tools for assessing and profiling structured datasets" and vocabularies to represent profiles in the past decades [4]. According to the study, the full range of available features may be categorized into seven groups: Qualitative, Provenance, Links, Licensing, Statistical, Dynamics, and Other. Part of our (RDF) graph-based *measures* (see Section 3) belongs to the group of Statistical features. However, most of the tools listed in the paper gather comprehensive statistics and summaries at instance- and/or schema-level, leaving out to target the topology.

In summary, the study of RDF-specific properties of publicly available RDF datasets has been extensively covered. It is currently supported by online services and tools, such as LODStats and Loupe. Therefore, in addition to these works, we focus on analyzing graph invariants in RDF datasets.

Graph-based Analyses

In the area of structural network analysis, it is common to study the distribution of specific graph measures in order to characterize a graph. RDF datasets and schemas have also been subject to these studies. Most of these works focus on studying different in- and out-degree distributions, path length, and are limited to one dataset or a rather small collection of RDF datasets, for instance, when investigating topological characteristics of one particular vocabulary of interest.

The study by Ding et al. [20] reveals that the power-law distribution at instance-level is prevalent across graph invariants in RDF graphs, obtained from 1.7 million documents. Theoharis et al. also investigated the schema level of RDF graphs [21]. Their study covers 250 schemata and concluded that the majority of classes with class descendants and property degree distributions approximate a power-law. Hu et al. studied entity links in the domain of Life Sciences [22] and discovered that the degree distribution of entity links does not strictly follow the power law.

The small-world phenomenon [23], known from experiments on social networks, were also studied within the Semantic Web [24, 25], with the result of saying that Linked Open Data is having the small-world characteristic [3]. Bachlechner et al. [25] found that the entire FOAF[5] network is a small-world with high local clustering coefficient and a power-law distribution. Their analysis showed that, in this network, the average degree is 9.56, with a diameter (characteristic path length) of 6.26. The work by Flores et al. [26] analyzes further relevant graph invariants in RDF graphs, such as statistics on the number of vertices and edges, in- and out-degree distributions, density, reciprocity, and *h*-index. The work by Flores et al. applied graph-based metrics on synthetic RDF datasets. More recently, Fernández et al. [3] have studied the structural features of real-world RDF data and the relatedness between vertices and edges in RDF graphs, using subject-object, subject-predicate, and predicate-object ratios. Their experimental study investigates fourteen real-world RDF datasets from seven categories, in order to find "common features and characterize real-world RDF data".

Complementary to these works, we present a study on 280 RDF datasets acquired from the LOD Cloud. We primarily focus on analyzing measure effectiveness and measure performance from a set of fifty-four graph-based measures. By this means, we will also get some understanding and insights into the structure of real-world RDF datasets.

3. Measures for RDF Graphs

In [2], we introduced a number of measures which are formalized here. The set of measures utilized in the experiments in the subsequent sections is complemented by the measures described and formalized by Fernández et al. in [3]. By this means, we can provide an understanding of their complementarity as a whole.

[5] http://xmlns.com/foaf/spec/

First, Section 3.1 introduces graph notations and definitions that are used throughout the paper. Section 3.2 then introduces definitions for all graph measures studied in [2]. Table 1 presents an overview of the graph measures described in this section.

3.1. Graph Data Model

Definition 3.1 (Directed Multigraph)**.** A *multigraph* G is a pair of finite sets (V, E), with V denoting the set of all vertices, and E a multiset of directed, labeled edges in the graph G.

In this work, for the sake of simplicity, we use the terms graph and multigraph interchangeably. They are used when referred to a *graph measure* or *graph invariant*. In particular, the RDF data model builds upon this definition to represent RDF graphs. RDF graphs [27] are multigraphs modeled as a set of RDF triples. RDF triples are composed of terms from U, B, L, which are disjoint finite sets of URI references, blank nodes, and RDF literals, respectively.

Definition 3.2 (RDF triple)**.** An *RDF triple* is a tuple $(s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)$. s is denoted as the *subject*, p the *predicate*, and o the *object*.

Through RDF triples, we can define RDF graphs [28].

Definition 3.3 (RDF graph)**.** An *RDF graph* G is a set of RDF triples, where each (s, p, o) becomes a directed labeled graph structure of the form $s \xrightarrow{p} o$.

The sets of subjects, predicates, and objects in the RDF graph G will be referred to as $S_G \subseteq (U \cup B)$, $P_G \subseteq U$, and $O_G \subseteq (U \cup B \cup L)$, respectively. When referring to the general graph topology V and E will denote the set of vertices and edges of the graph G. Moreover, with respect to the RDF terminology, V is the set of all subjects and objects, i.e., $V = \{v \mid v \in (S_G \cup O_G)\}$. Note that, the set of vertices V may also contain predicates, as predicates are subjects within the schema-definition (TBox, if defined), and therefore elements of S_G. As given in the definition above, E is a multiset of (labeled) edges, since a pair of subject and object resources may be described with multiple RDF predicates. For example, in the graph $\{s\ p1\ o\ .\ s\ p2\ o\}$, E has two pairs of vertices, and therefore $E = \{(s,o)_1, (s,o)_2\}$.

3.2. Graph Measures

3.2.1. Basic Graph Measures

In the following, we describe measures that can be applied to graphs in general (cf. Definition 3.1).

We report on the total **number of vertices** n and the total **number of edges** m for a graph G. Some works in the literature refer to these values as size and volume, respectively. These measures are relevant, as the number of vertices and edges usually varies drastically across knowledge domains.

$$n = |V| \tag{1}$$

$$m = |E| \tag{2}$$

In multigraphs, parallel edges represent edges that share the same pair of source and target vertices. Therefore, the measure **number of parallel edges**, denoted as m_p, is defined as

$$m_p = |\{e \mid count_e(e, E) > 1, e \in E\}| \tag{3}$$

with $count_e(e, E)$ being a function that returns the multiplicity of e in E, i.e., number of times e is contained in E. Based on the above measure, we also compute the total number of edges without counting parallel edges, called the **number of unique edges**, denoted as m_u. This measure will give us an impression of the "raw" shape of the graph, which is useful when one may want to study graph clustering, like in a network, for instance. It is computed by subtracting m_p from the total number of edges m, i.e.

$$m_u = m - m_p \tag{4}$$

3.2.2. Degree-based Measures

In a graph $G = (V, E)$, the **degree** of a vertex $v \in V$ is the total number of edges that are connected to it. With directed graphs, as is the case of RDF graphs, it is common to distinguish between **in-degree** and **out-degree** of a vertex v. For a given $v \in V$, we define the total degree by means of the in- and out-degree.

$$d(v) = d^+(v) + d^-(v) \tag{5}$$

with

$$d^+(v) = |\{(u, v) \mid \exists\, u \in V, (u, v) \in E\}| \tag{6}$$

$$d^-(v) = |\{(v, u) \mid \exists\, u \in V, (v, u) \in E\}| \tag{7}$$

The previous definitions of d^+ and d^- also take into account parallel edges.

In social network analyses, vertices with a high out-degree are said to be "influential", whereas vertices with a high in-degree are called "prestigious". To identify these vertices in an RDF graph, we compute the **maximum total-, in-, and out-degree** of the graph's vertices, denoted as $d = \max_{v \in V} d(v)$, $d^+ = \max_{v \in V} d^+(v)$, $d^- = \max_{v \in V} d^-(v)$, respectively. In addition, we compute the graph's **mean total-, in-, and out-degree** denoted z, z^+, and z^-, respectively.

These measures may be applied in research about RDF data management, for instance, where the (average) degree of a vertex (database table record) has a significant impact on query evaluation, since queries on dense graphs can be more costly in terms of execution time [29].

Another degree-based measure is ***h*-index**, known from citation networks [30]. In a graph G a value of h means that for the number of h vertices in the graph, the degree of these vertices is greater or equal to h. In order to compute the value through the following equation, as a prerequisite, it is required to have a list of all vertex degrees sorted in descending order.

$$h = \max_{i \in |V|} \min\ (d(v_i), i), v_i \in V \tag{8}$$

with i being the position in the list and $d(v_i)$ the degree of the vertex at the i-th position.

This measure may be an indicator of the importance of a vertex, similar to a centrality measure (see Section 3.2.3). Further, a high value of a graph's $h-$index could be an indicator for a "dense" graph and that its vertices are more "prestigious". In this work, we report on this network measure for the directed graph (using only the in-degree of vertices) denoted as h^+ and the undirected graph (using in- and out-degree of vertices) denoted as h.

3.2.3. Centrality Measures

In social network analyses, the concept of *point centrality* expresses the importance of nodes in a network. There are many interpretations for the term "importance" and so are measures for centrality [1]. A high centrality value of a vertex generally means that it is more "important", although for different reasons, as indicated by the different measures.

Using the degree of a vertex, $d(v)$, point centrality is denoted as C_d. To indicate that it is a centrality measure, and not just the degree, the literature often normalizes these values by the total number of all

Table 1

Set of graph measures implemented and evaluated in this study

Measure Name	Value	Symbol	Measure Group	Comment
vertices	max	n	basic	-
edges	max	m	basic	-
parallel edges	max	m_p	basic	-
unique edges	max	m_u	basic	-
total degree	max | mean	$d \mid z$	degree-based	-
in-degree	max | mean	$d^+ \mid z^+$	degree-based	-
out-degree	max | mean	$d^- \mid z^-$	degree-based	-
h-index directed	-	h^+	degree-based	Employing the in-degree of the vertices.
h-index undirected	-	h	degree-based	Employing the total-degree of the vertices.
degree centrality	max	C_d	centrality	-
in-degree centrality	max	C_{d^+}	centrality	-
out-degree centrality	max	C_{d^-}	centrality	-
centralization degree	-	C_d^+	centrality	-
page-rank	max	PR	centrality	-
fill overall	max	p	edge-based	Respects all edges, i.e. including parallel edges.
fill unique	max	p_u	edge-based	Respects only unique edges.
reciprocity	max	y	edge-based	-
diameter	max	δ	edge-based	Approximated value using pseudo-diameter algorithm[6].
variance in-degree	-	σ^{2+}	descriptive stat.	-
variance out-degree	-	σ^{2-}	descriptive stat.	-
std.dev. in-degree	-	σ^+	descriptive stat.	-
std.dev. out-degree	-	σ^-	descriptive stat.	-
coeff.variation in-degree	-	cv^+	descriptive stat.	-
coeff.variation out-degree	-	cv^-	descriptive stat.	-
degree powerlaw exp.	-	α	descriptive stat.	-
in-degree powerlaw exp.	-	α^+	descriptive stat.	-

vertices. We compute the **maximum point centrality** for the graph G as

$$C_d = \frac{d}{n}, \text{ with } d = \max_{v \in V} d(v) \tag{9}$$

Besides the point centrality, there is also the measure of *graph centralization* [31], which is known from social network analysis. This measure may also be seen as an indicator of the type of graph. It expresses the degree of inequality and concentration of vertices by means of a perfectly star-shaped graph, which itself is at most centralized and unequal with regard to its degree distribution. The **graph centralization** value of one graph G regarding the degree is defined as:

$$C_d^+ = \frac{\sum^{v \in V}(d - d(v))}{(n-1)*(n-2)} \tag{10}$$

Another centrality measure is PageRank [32], which considers all incoming edges to a vertex to estimate its importance. After computing the PageRank value for all vertices $v \in V$ in the graph G, denoted as $PR(V)$, the **maximum PageRank** value is defined as

$$PR = \max_{v \in |V|} PR(v) \tag{11}$$

3.2.4. *Edge-based Measures*

As the (average) number of vertices and edges vary highly across knowledge domains [2], it is interesting to measure the so-called "density" of a graph, sometimes referred to as "connectance" or "fill". The density is computed as the ratio of all edges to the total number of all possible edges. The formula is in accordance with the definition of RDF graphs, which are directed and may contain loops. As mentioned earlier, RDF graphs may contain parallel edges, and thus we provide an additional measure, which uses unique edges only. Therefore, **fill_overall** and **fill**, denoted as p and p_u, respectively, are defined as follows:

$$p = \frac{m}{n^2} \tag{12}$$

$$p_u = \frac{m_u}{n^2} \tag{13}$$

These measures may be used to calculate the probability of an edge between two randomly chosen vertices in the graph G. Comparing the measure fill with centrality measures shows that dense graphs show higher centrality values of the vertices, which in turn leads to higher "connectivity" and linkage among them, as mentioned earlier. This also has a positive impact on navigation through the graph.

As RDF graphs are directed and labeled graphs, the aspect of "navigability" through the graph through RDF predicates is of interest. We analyze the fraction of bidirectional connections between vertices in the graph. These are pairs of vertices forward- connected by some edge, which are also backward-connected by some other edge. The value of **reciprocity**, denoted as y, is expressed as the ratio of the **number of bidirectional edges**, denoted as m_{bi}, among all edges in the graph G

$$y = \frac{m_{bi}}{m} \tag{14}$$

with

$$m_{bi} = |\{(u, v) \in E \mid \exists (v, u) \in E\}| \tag{15}$$

High values of reciprocity mean there are many links between vertices that are bidirectional. This value is typically high in citation or social networks.

Another critical group of measures that is described by the graph topology is related to paths. A path is a set of edges one can follow along between two vertices. As there can be more than one path, the **diameter** is defined as the longest shortest path between two vertices of the network [1], denoted as δ.

$$\delta = \max_{v,u \in V} path(v, u) \tag{16}$$

The diameter is usually a very time-consuming measure to compute since all possible paths have to be considered. Thus, we used the `pseudo diameter` algorithm[6] to estimate the value of the diameter for the studied RDF graphs. In query optimization over RDF data, this measure may be applied to estimate the cardinality of joins (e.g., subject-object joins), which heavily depends on the paths in an RDF graph.

3.2.5. Descriptive Statistical Measures

Descriptive statistical measures are useful to describe distributions of some set of values. It can be useful to consult the **degree of dispersion** of the distribution of interest; in our scenario, it is the distribution of vertex degrees in the graphs. Types of dispersion are, for example, the **degree variance** σ^2, and the **degree standard deviation** σ,

$$\sigma^2 = \frac{\sum_{v \in V}(d(v) - z)^2}{n - 1} \tag{17}$$

$$\sigma = \sqrt{\sigma^2} \tag{18}$$

We compute these measures also for the in- and out-degree distributions of vertices in the graphs, denoted as σ^{2+}, σ^{2-}, and σ^+, σ^-, respectively. They are defined adequately using the appropriate in- and out-degree values for vertex degree and mean degree of all vertices V of a graph.

However, when one would want to compare different standard deviation values, it would not be very meaningful, since they most probably are computed using different means. The **coefficient of variation**, denoted as cv, may be consulted to have a comparable measure for distributions with different mean values. It is obtained by dividing the standard deviation σ by the corresponding mean z.

$$cv = \frac{\sigma}{z} \tag{19}$$

cv may also report the type of distribution concerning a set of values. For example, a low value of cv^- means a constant influence of vertices in the graph (homogeneous group). In contrast, a high value of cv^+ means high prominence of some vertices in the graph (heterogeneous group).

Further, the type of *degree* distribution is an often considered measure of graphs. In some knowledge domains, datasets report on degree distributions that follow a power-law function [22], which means that the number of vertices with degree k behaves proportionally to the power of $k^{-\alpha}$, for some $\alpha \in \mathbb{R}$.

[6] https://graph-tool.skewed.de/static/doc/topology.html#graph_tool.topology.pseudo_diameter

Such networks are called scale-free. The literature has found that values in the range of $2 < \alpha < 3$ are typical in many real-world networks [1]. The scale-free behavior also applies to some datasets and measures of RDF datasets [3, 20]. However, to reason about whether a distribution follows a power-law can be technically challenging [33], and computing the exponent α, that falls into a specific range of values, is not sufficient. We compute the exponent for the total- and in-degree distributions [33], denoted as α and α_{in}, respectively. Also, to support the analysis of power-law distributions, the framework produces plots for both distributions. A power-law distribution is described as a line in a log-log plot.

Determining the function that fits the distribution may be of high value to estimate the selectivity of vertices and attributes in graphs. The structure and size of datasets created by synthetic datasets, for instance, can be controlled with these measures. Also, an explicit power-law distribution allows for high compression rates of RDF datasets [3].

4. Performance of Graph Measures for Dataset Profiling - Research Questions and Setup

Building on the implementations of graph measures introduced in the previous section, this section introduces an experimental investigation into the performance of measures for describing, profiling, and distinguishing datasets. Whereas Section 4.1 presents our research questions and motivates the experiments, Section 4.2 describes the design and methodology of the experiments which apply and assess our measures on datasets from the LOD Cloud through established feature selection and analysis techniques.

4.1. Research Questions

This section elaborates on the research questions which motivated our experiment. Let M denote the set of all measures employed in our experiments. Further, let C denote the set of all knowledge domains, i.e., categories or classes, available in the LOD-Cloud. D_c denotes the set of datasets assigned to the corresponding category $c \in C$.

A (graph) measure is a feature in the context of statistical operations (correlations, feature engineering, statistical learning algorithms). Starting from here, we will use these terms interchangeably. The usage of the corresponding terms should be clear from the context.

RQ1: What is an efficient and non-redundant set of features for characterizing RDF graphs?

In order to characterize graphs or sets of graphs within domains efficiently, concise graph descriptions have to be based on efficient, non-redundant feature sets where each feature provides significant information gain.

This question aims at finding a concise and finite set $M' \subset M$ of measures that reduce or eliminate redundancy and maximize information gain through correlation analysis. This step will improve the effectiveness of the resulting set of graph measures and improve their applicability, for instance, as part of machine learning models.

RQ2: Which measures describe and characterize individual knowledge domains most/least efficiently?

Datasets within the LOD cloud are categorized into nine distinct knowledge domains so that each dataset is associated with precisely one specific category. In order to understand the representativeness and variability of topological measures within a knowledge domain, we investigate the heterogeneity of feature values within and across distinct domains through basic statistic metrics and discuss observed values representative for distinct LOD domains. We will refer to this feature set as M''_c with $c \in C$. Please note that $M''_c \subset M', \forall c \in C$.

This will provide insights into the capacity of individual features to represent the nature of particular domains and may contribute to discriminative models and to filtering out noise features when profiling datasets.

RQ3: Which measures show the best performance to discriminate individual knowledge domains?

Datasets from a knowledge domain exhibit distinct characteristics with respect to topological features of the graphs but also with respect to other features, such as vocabulary adoption. A particular question is which (RDF-) graph measures are most descriptive *within* one particular knowledge domain. In contrast to RQ2, this research question investigates feature importance for each domain. The findings are of interest to synthetic dataset generators, for example. By generating a synthetic dataset, benchmark suites most often target some particular domain of interest. When generating datasets for the *Publications* knowledge domain, for example, a generator should

Table 2

Statistics on RDF datasets which were acquired for the experiments. Listed are the number of RDF datasets per knowledge domain and their corresponding maximum and average number of vertices *n* and edges *m*

Domain	**# datasets**	**Maximum**		**Average**	
		n	***m***	***n***	***m***
Cross Domain	15	291,178,702	1,042,217,722	36,276,052	111,329,448
Geography	11	47,541,174	340,880,391	9,763,721	61,049,429
Government	37	131,634,287	1,489,689,235	7,491,531	71,263,878
Life Sciences	32	356,837,444	722,889,087	25,550,646	85,262,882
Linguistics	122	120,683,397	291,314,466	1,260,455	3,347,268
Media	6	48,318,259	161,749,815	9,504,622	31,100,859
Publications	50	218,757,266	720,668,819	9,036,204	28,017,502
Social Networking	3	331,647	1,600,499	237,003	1,062,986
User Generated	4	2,961,628	4,932,352	967,798	1,992,069

follow a specific set of measures, range of values, and used vocabularies, in order to be identified with that category of datasets.

4.2. Experimental Setup

Section 4.2.1 explains which datasets were acquired and used for our experiment. Section 4.2.2 gives details about the framework and the measure computation. Section 4.2.3 explains how measure efficiency and measure importance were obtained.

4.2.1. Datasets

We have downloaded a large group of datasets from the LOD Cloud 2017[7] and prepared it with our framework presented in [2].

From the total number of 1,163 potentially available datasets in the LOD Cloud 2017, 280 datasets were selected based on the criteria: (i) RDF media types statements that were correct for the datasets, and (ii) the availability of data dumps provided by the services. To not stress SPARQL endpoints to transfer large amounts of data, in this experiment, only datasets that provide downloadable dumps were considered.

To dereference RDF datasets, we relied on the metadata (so called data-package) available at DataHub, which specifies URLs and media types for the corresponding data provider of one dataset[8]. We collected the datapackage metadata for all datasets and manually mapped the obtained media types from the datapackage to their corresponding official media types that are given in the specifications. For instance, `rdf`, `xml_rdf` or `rdf_xml` were mapped to `application/rdf+xml` and similar.[9] In this way, we obtained the URLs of 890 RDF datasets. After that, we checked whether the dumps are available by performing `HTTP HEAD` requests on the URLs. At the time of the experiment, this returned 486 potential RDF dataset dumps to download. For the other not available URLs, we verified the status of those datasets with http://stats.lod2.eu. After these manual preparation steps, the data dumps could be downloaded with the framework.

The framework needs to transform all formats into N-Triples. From here, the number of prepared datasets for the analysis further reduced to 280. The reasons were: (1) corrupt downloads, (2) wrong file media type statements, and (3) syntax errors or other formats than these what were expected during the transformation process. This number seems low compared to the total number of available datasets in the LOD Cloud, though it sounds reasonable compared to recent studies on the LOD Cloud [17–19]. Table 2 gives some descriptive statistics about the analyzed datasets.

As graph library we used *graph-tool*[10], an efficient library for statistical analysis of graphs. In *graph-tool*, core data structures and algorithms are implemented in C/C^{++}, while the library itself can be used with Python. *graph-tool* comes with pre-defined implementations for graph analysis, e.g., degree distributions or more advanced implementations on graphs like PageRank or clustering coefficient. Further, some values may be stored as attributes of

[7] http://lod-cloud.net/versions/2017-08-22/datasets_22-08-2017.tsv.

[8] Example: https://old.datahub.io/dataset/<dataset-name>/datapackage.json

[9] Other media type statements like `html_json_ld_ttl_rdf_xml` or `rdf_xml_turtle_html` were ignored, since they are ambiguous.

[10] graph-tool, https://graph-tool.skewed.de/

Table 3
Set of twenty-nine RDF graph measures, which were implemented and evaluated in this study

Measure Name	Value	Group
out-degree	max | mean	subject out-degrees
partial out-degree	max | mean	subject out-degrees
labelled out-degree	max | mean	subject out-degrees
direct out-degree	max | mean	subject out-degrees
in-degree	max | mean	object in-degrees
partial in-degree	max | mean	object in-degrees
labelled in-degree	max | mean	object in-degrees
direct in-degree	max | mean	object in-degrees
subject/object ratio	ratio	common ratios
degree	max | mean	predicate degree
in-degree	max | mean	predicate degree
out-degree	max | mean	predicate degree
repeated predicate list	ratio	predicate lists
predicate list degree	max | mean	predicate lists
distinct classes	max	typed subjects/objects
typed subjects	max	typed subjects/objects
ratio of typed subjects	ratio	typed subjects/objects

vertices or edges in the graph structure. The library's internal graph-structure may be serialized as a compressed binary object for future re-use. It can be reloaded by *graph-tool* with much higher performance than the original edgelist. We instantiated the graphs from the binary representation (see next section) and operated on the graph objects provided by the *graph-tool* library.

4.2.2. Graph Measures Computation

All graph-based measures introduced in Section 3.2 where already part of the framework introduced in [2]. In order to do a more comprehensive evaluation of the effectiveness of graph measures, we include RDF graph measures from Fernández et al. [3], who provides a comprehensive list and formalization of various RDF graph-based measures. Table 3 gives an overview of all RDF graph-measures we implemented as a module extension[3] of our framework. In order to optimize performance, we worked with lists of vertices, edges, and edge labels (predicates), using Python's build-in operations for lists and additional libraries for scientific computing in Python, like *numpy*[11], and *pandas*[12]. That way, the computation of measures, such as the maximum and mean in-/out-degree of all vertices, was straight-forward. A more complex example is the partial out-degree measure, which is "defined as the number of triples of G in which s occurs as subject and p as predicate". In order to compute this measure from the perspective of a native graph object in memory, one must create an array of all pairs of source vertices (subjects) and their outgoing edge labels (predicates) and count the number of grouped occurrences of these pairs.

On graphs with a particularly large number of edges ($> 100,000,000$) the building temporary lists of edge labels and the linear iteration over lists of vertices is not of acceptable performance. Therefore, we employed graph partitioning mechanisms for a large number of measures, in order to compute the desired values in a *map-reduce*-fashion. We encourage the interested reader to look into the corresponding package of the framework[3] to find the implementation for all measures.

4.2.3. Measure Efficiency and Measure Importance

For RQ1, we will first give an overview of all the measures and their relationship among each other by calculating the Spearman correlation coefficients between all measures. To this end, the Spearman correlation test is employed, since most of the distributions of measure values do not follow a normal distribution. To reduce the number of measures, we employ two popular methods: (a) a low variance test, which filters measures which fall below a certain threshold, and (b) popular univariate statistical tests, from which we choose Chi2, and Mutual Information (MI). Since many of the variables are continuous, and MI only works with discrete values, Maximum Information Non-parametric Estimation (MINE) is utilized additionally. Therefore, M' is defined as follows:

$$M' = \{m \in M \mid threshold(m, F) \geqslant 3\}, \qquad (20)$$

with F being the set of all feature selection methods mentioned above. $threshold()$ returns the number of methods having a match over the given measure m.

For RQ2, we will show boxplots as aggregated descriptive statistics for some selected measures. This will give insights into the distribution of values. In order to investigate the variability at the category level, we apply some statistical methods. To show the variability *per category*, we group all datasets by categories and compute the variance per measure and

[11] numpy, the fundamental package for scientific computing with Python, https://numpy.org/

[12] pandas, a library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language, https://pandas.pydata.org/

group. By this means, we can analyze noisy and non-noisy features in terms of variance and assign the corresponding M''_c for all $c \in C$.

The *variability across categories* (*vac*) is computed by taking the mean of a measure for all datasets in a particular category $c \in C$ computing the standard deviation over the obtained means subsequently. More formally, with *val* denoting all values for measure $m \in M'$ and datasets in D_c, with $c \in C$

$$vac(m) = std(\{\ \overline{val(m, D_c)} \mid c \in C\ \}) \tag{21}$$

For the classification tasks in RQ3, we deploy and tune a Random Forest classifier for both tasks. Initial experiments have shown that Random Forest outperforms other established classification on our task. Measure efficiency/performance is evaluated in two different experiments. First, we will train a classifier in order to predict one of all six domains. By means of this classification task, we will investigate measure performance, in order to discriminate all domains between each other. Second, in another classification task, we want to find those measures with the best performance to describe one particular knowledge domain. This is done by employing the binary relevance method, which is a one-vs-rest version of the first classification task. It will evaluate measure performance for each individual domain by training one independent classifier per domain. The measures with the best performance will have the ability to characterize datasets within one particular category most effectively.

Please note that our main aim is to understand overall and class-wise feature (i.e., graph measure) importance, rather than finding the best model for predicting category labels of RDF graphs. However, we want to find meaningful results. Thus we are obliged to tune the classifier to some extend. We hyper-tune the parameters via grid-search and five-fold cross-validation.

Since the classes are not balanced (cf. Table 2), we experimented with over- and undersampling strategies. For oversampling, we used the SMOTE-algorithm, for undersampling, a random undersampler. The results are presented by employing the highest scored classifier from the parameter-tuning and sampling strategy.

4.3. Execution Environment

The operating system, client software, database (with the records for all measures), reside all on one server during the experiment. The experiments were performed on a rack server Dell PowerBridge R720, having two Intel(R) Xeon(R) E5-2600 processors with 16 cores each, 192GB of main memory, and a 10TB total main storage. The operating system was Ubuntu 18.04.1 LTS, kernel version 4.15. Docker image version with the corresponding *graph-tool*[10] library was 2.29. All RDF graph measures shown in Table 3 were computed directly on the instantiated graph-object after loading into memory.

The computation of the measures on the graphs requires much physical memory. For graphs with less than 100M edges, the framework was configured to work in parallel with 12 concurrent processes. All other graphs (more than 100M edges) were computed one after another.

5. Assessing Graph Measures of the Linked Open Data Cloud - Results

We present our results by referring to the research questions. A more detailed discussion about the results can be found in the follow-up section (cf. Section 6).

5.1. RQ1: What is an efficient and non-redundant set of features for characterizing RDF graphs?

Correlation coefficients

We first report on observations about correlation coefficients between measures. Figure 1 shows a correlation matrix of all measures, encoded as blue (strong positive correlation), light (no correlation), to red (strong negative correlation). The values were computed with the Spearman correlation test.

In the group of graph measures, the number of edges `m` and vertices `n` has an almost perfect correlation with (a) `max_degree` and (b) `max_in_degree`. In addition, the two measures have a strong positive mutual correlation. Due to this, other measures which employ these measures are in turn strongly correlated with each other. In particular, this can be observed for measures employing the in-degree. Descriptive statistics on the distribution of in-degrees, like `var_in_degree`, `stddev_in_degree`, and `coefficient_var_in_degree`, grow with the size and volume of the graphs. This does not apply for measures relating to the vertices out-degree: measures using the in-degree differ from measures using the out-degree. Most of the measures employing the out-degree do not correlate with almost any of the

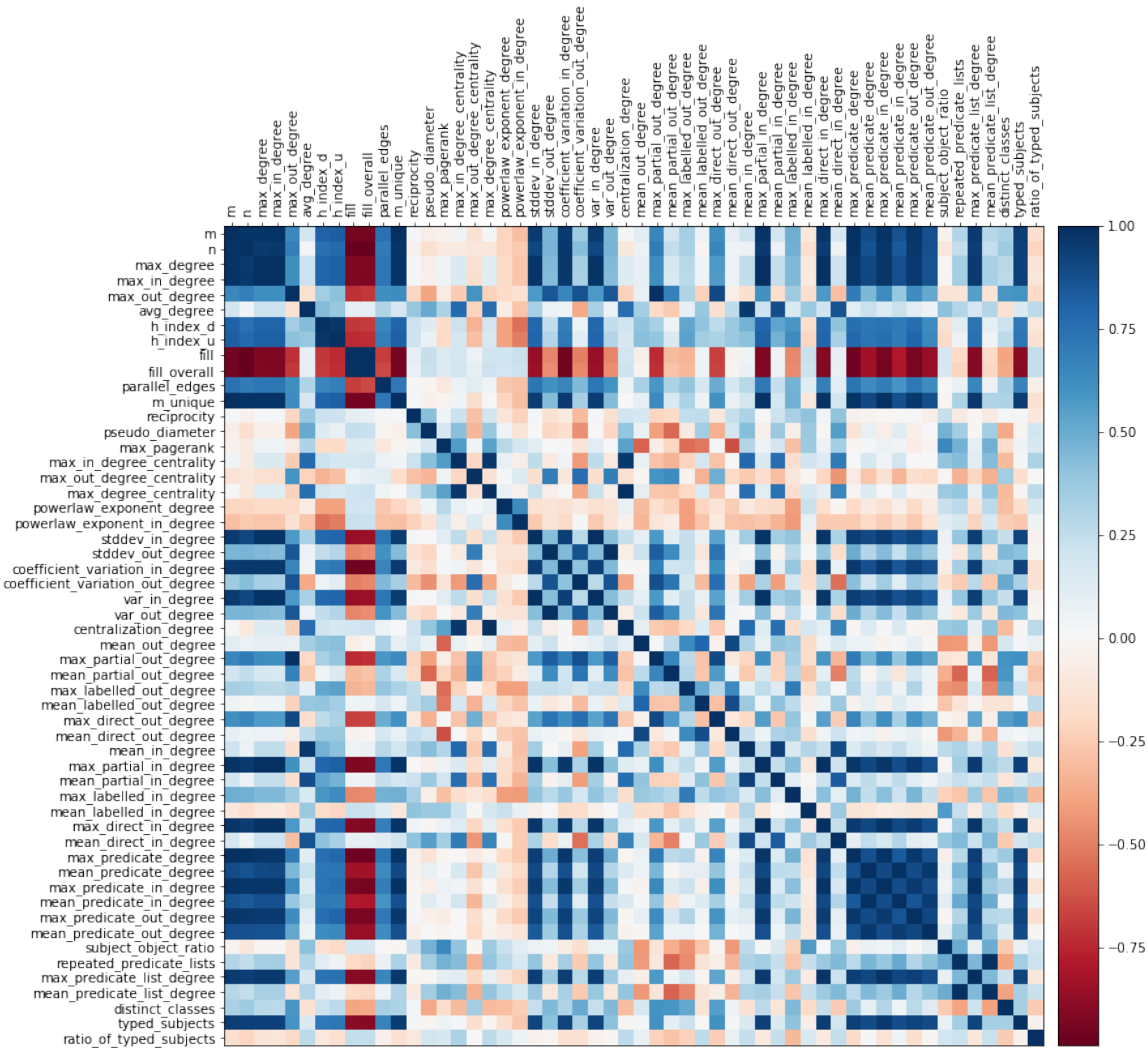

Fig. 1. **Correlation matrix**. Shows the whole set of measures and their Spearman correlation coefficient encoded as blue (strong positive correlation), light (no correlation), to red (strong negative correlation).

other measures, which makes them more descriptive. A negative correlation value implies that while values for a measure *x* increase, values for another measure *y* decrease. This is the case with measures employing the aspect of density (`fill`) of the graphs with increasing size `n` and volume `m`. The density of a graph also has a negative correlation to the distribution of vertex degrees, as we can see with variance, standard deviation, and coefficient of variation values. This means that the denser the graphs are (`fill` increases), the more homogeneous the vertex degrees of the graphs become (descriptive statistics over vertex degrees become smaller). Almost no dependencies are exhibited by `avg_degree`, `reciprocity`, `diameter`, centrality measures, and the `powerlaw_exponent`s, which measures the type of distribution of vertex (in-)degrees.

In the group of RDF graph measures, there are less inter-relationships. As a group, measures employing predicate degrees, `max_predicate_list_degree`, together with `max_partial_in_degree`, `max_direct_in_degree` as well as the `typed_subjects` measure, have strong positive mutual correlations. All of the mentioned measures

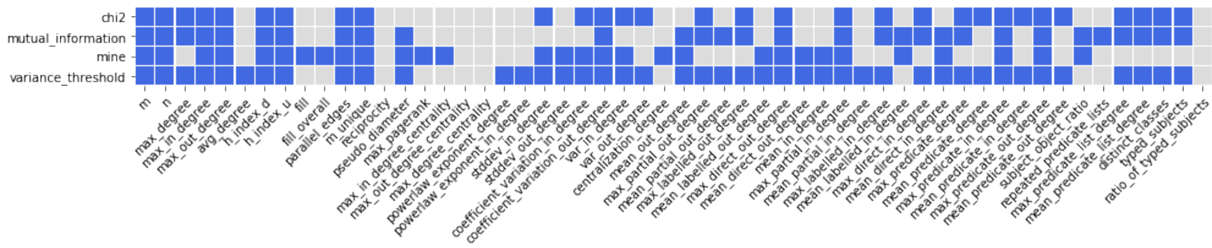

Fig. 2. **Meaningful measures** (highlighted in blue) according to different statistical feature selection scoring methods.

grow with the size `n` and volume `m` of the graphs. Some individual mutual strong positive correlations can be observed, for instance, between `repeated_predicate_lists` and `mean_predicate_list_degree`, `mean_direct_in_degree` and `mean_in_degree` and `mean_partial_in_degree`. As in the first group of graph measures, all "mean" in-degree measures have strong correlations among each other as well as to the `mean_degree`.

Measure selection

Figure 2 highlights in blue the measures that were selected by the individual tests.

Overall, there is variance and no particular consensus of the statistical tests. However, there are some agreements. Looking at agreements in *all* tests, only thirteen measures are providing information gain; only three were dismissed by all tests, i.e., two degree-centrality measures and `ratio_of_typed_subjects`. Sixteen measures have agreements in three tests (threshold was not met in one of the tests); ten measures met the threshold in only one test. With thirty measures, the pair Chi2 and Variance Threshold has the highest number of agreements; Mutual Information and Variance Threshold agree on twenty-seven measures. The least agreements can be found for the pair Mutual Information and MINE (eighteen).

Summary of results

With particular regard to RDF graphs and the above analysis, we conclude with the following observations:

- The larger the density, the more "stable" and homogeneous is the (in-/out-) degree distribution of vertices in the graphs.
- The larger the size and volume of the graphs, the more typed subjects become present, and the higher the number of subjects using a fixed set of predicates appears (cf. predicate degree and predicate lists measures).
- The average degree of the graphs is mainly influenced by the in-degree.
- Measures employing the distribution of out-degrees are more descriptive.

The next subsections report the results on the reduced set of meaningful measures obtained from the feature selection methods. In particular, M' is defined as the set of measures where at least three of the tests have an agreement, i.e., $|M'| = 29$.

5.2. RQ2: Which measures and values describe and characterize knowledge domains most/least efficiently?

In order to get a sense of the variability of measures within and across knowledge domains, in this section, we look closer and report on characteristics for some individual measures first. Afterwards, we aggregate and report on variability across knowledge domains, through variance and standard deviation.

Characteristics of values

Figure 3 shows, by example, the distribution of values for two groups of measures. The first group at the top row shows exemplary measures which were sorted out by the feature selection approaches in Section 5.1, such as the mean total-degree and the mean out-degree; the bottom row shows exemplary features of M'. The figure shows all available knowledge domains except Media, Social Networking, and the User Generated, due to few dataset retrieved in these categories (cf. Table 2).

Regarding the mean total-degree, some categories show very similar median values, like *Cross Domain*, *Life Sciences*, and *Publications*. *Cross Domain*, *Geography*, *Life Sciences*, and *Linguistics* share a similar maximum value. However, regarding the outliers, *Life Sciences* contains a dataset that has by far the highest average degree, followed by a dataset in the *Government* category. The mean out-degree

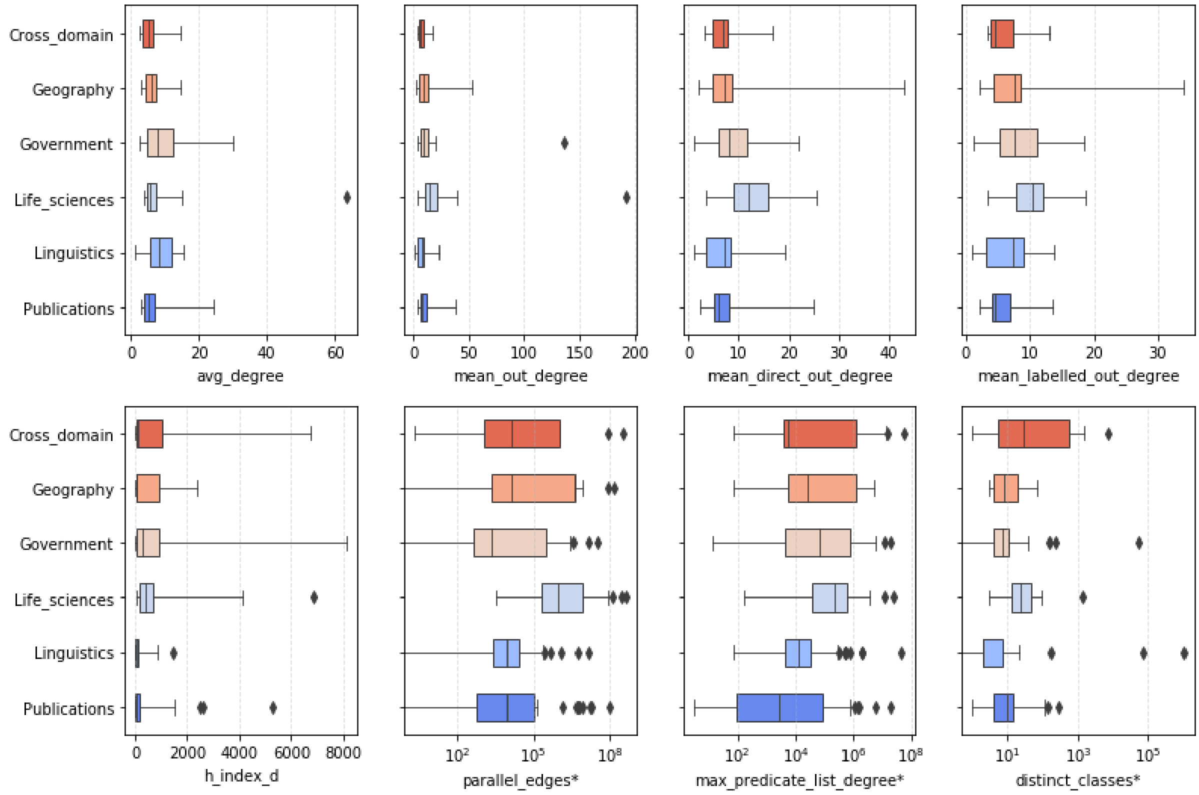

Fig. 3. **Descriptive statistics** of measures sorted out by feature selection methods (top) and measures considered meaningful (bottom). * indicates that x-axis is log-scaled

(outgoing predicates of subjects) is higher for most of the categories (two outliers can be observed with very high values). The boxes reveal that the majority of values are larger than the mean total-degree, which means that the mean total-degree is mainly influenced by the in-degree. This is particularly striking for datasets in the *Geography* and *Life Sciences* domains.

The last two plots in the first group show the `mean_direct_out_degree` and `mean_labelled_out_degree` measures, which describe the relationship of subjects to their average number of different objects and predicates, respectively. Overall, the number of different objects is higher than the number of predicates. The distribution of values is similar for *Cross Domain* and *Publications*, as well as for *Geography* and *Linguistics*, particularly for the predicates (`mean_labelled_out_degree`). Comparing `mean_degree` and `mean_out_degree` as well as `mean_direct_out_degree` and `mean_labelled_out_degree` with each other, we can see that they show very similar characteristics. Generally, the distribution of values is not symmetric (different whisker lengths of the boxes) and skewed, thus they do not follow a normal distribution. Further, there is little variability (short length of boxes).

Below in Figure 3 are exemplary measures of M', i.e., those that were considered to be non-redundant and meaningful according to the feature selection approaches in Section 5.1. There is much higher variability in most of the measures and knowledge domains. Also, the number of outliers is larger. Please note that the x-axis is log scaled for some measures, which makes it hard to make statements about the skewness of the distributions; thus, we would like to point out `h_index_d`. It gives us the number of at least `x` RDF objects with `x` incoming predicates.

Lowest spread and little variability can be found for `h_index_d`. The distribution of values in *Cross Domain*, *Geography*, *Government* is highly skewed to the right, which means that most of the values are rather low. However, there are some datasets with quite high value above 4000, e.g., in *Cross Domain*, *Government*, *Life Sciences*, and *Publications*. The

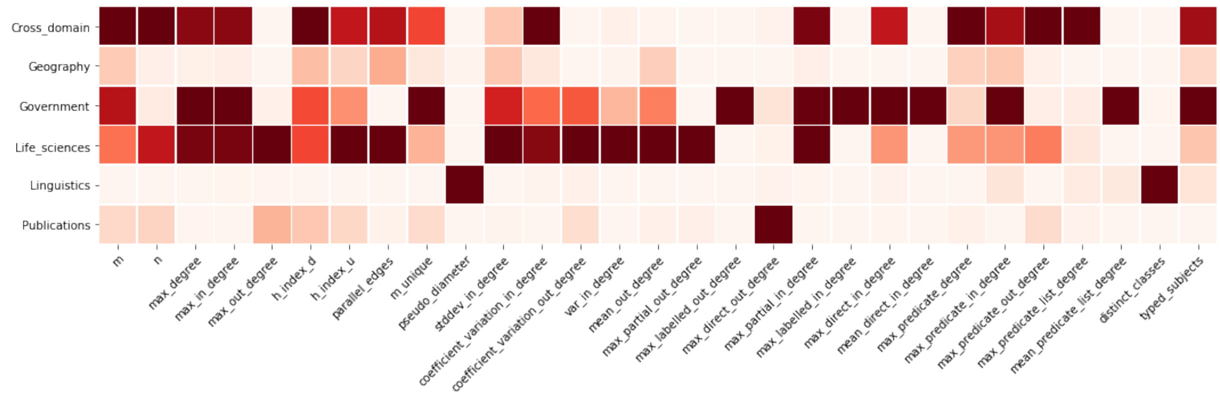

Fig. 4. **Measure variance**. The lighter the color the lower the variance and the more homogeneous the values are within the corresponding category.

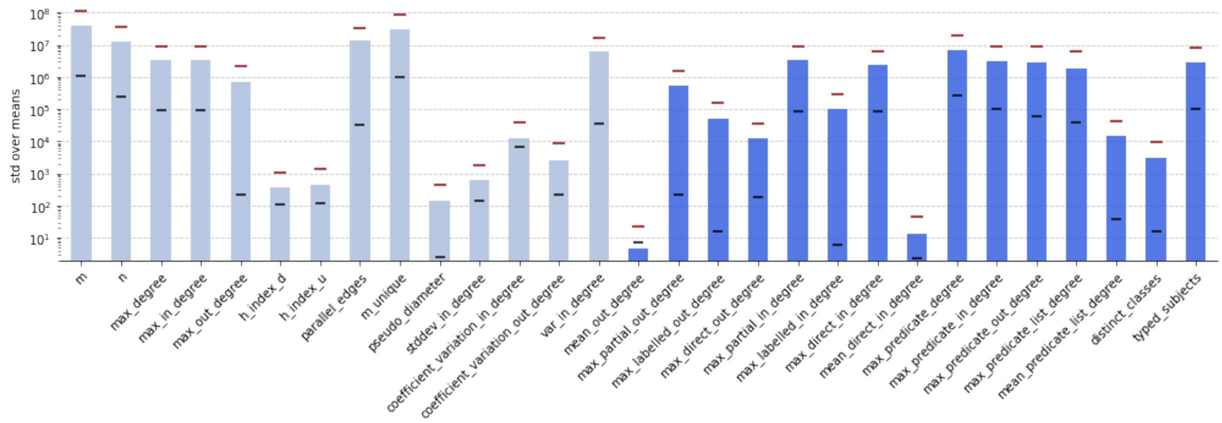

Fig. 5. **Degree of variance across knowledge domains**. A low/high value indicates low/high variance across knowledge domains. Colors encode graph (in light-blue) and RDF graph measures (in blue). y-axis is log-scaled.

largest value can be found for a dataset in the *Government* domain.

Variability of values

As a first overview, Figure 4 shows measure variance of the datasets within the given categories as a heat-map: the lighter the color, the lower the variance and therefore the more homogeneous the corresponding values are for the corresponding category and measure.

Overall, datasets in the *Life Sciences*, *Cross Domain*, and *Government* (in this order) have quite heterogeneous distributions of values for a high number of measures. On the contrary, only one, two, or three measures have high variance in the *Publications*, *Linguistics*, and *Geography* domain (in this order). Some measures exhibit high variance in just one category and a low variance in the others. Just to name a few: `max_out_degree` and `max_partial_out_degree` in *Life Sciences*, `pseudo_diameter` and `distinct_classes` in *Linguistics*, `max_labelled_in_degree` and `mean_predicate_list_degree` in *Government*, `max_predicate_list_degree` in *Cross Domain*, `max_direct_out_degree` in *Publications*. These measures may be used to discriminate categories against each other very well, as their characteristically distribution of values for a particular category can be considered meaningful. In turn, some measures also exhibit a rather low variance in one or two domains and higher in the others. These are, for instance, `m`, `h_index_d`, `std_in_degree` in *Linguistics*.

Figure 5 shows the degree of variance across knowledge domains. The scores are obtained by grouping datasets by category, taking the mean of the corresponding measure for all datasets per category, and then computing the standard deviation over these means. Lowest variances across all categories can be found for `mean_out_degree`, `mean_direct_in_degree`, `pseudo_diameter`, both `h_index` measures, `std_dev_in_degree`, `coefficient_variation_out_degree`, `distinct_classes`, and `mean_predicate_list_degree`. Among the top five measures with large dispersion between categories (`m`, `m_unique`, `parallel_edges`, `n`, and `max_predicate_degree`) are four measures employing graph edges. The figure also includes minimum (dark blue) and maximum (red) values. For some measures, the minimum value varies significantly from the standard deviation value. To name a few: `pseudo_diameter`, `max_labelled_in_degree`, `max_predicate_list_degree`, and `distinct_classes`.

Summary of results

- For the majority of the measures, the distribution of values is not normally distributed.
- The degree of variance across domains is significant for most of the measures. A low variance across domains is rather exceptional.
- Datasets in *Cross Domain* are heterogeneous, i.e., largest variability of the number of classes. While individual datasets have a high number of distinct classes, the variability within categories is less significant. Additionally, the number of typed subjects highly varies.
- Datasets in the *Government* domain have high variance in the mean degree of predicate lists, meaning that they are not homogeneous in terms of the used predicates per subject.
- Datasets in the *Linguistics* domain have high diameter[6].
- Each knowledge domain has datasets (graphs) with unique characteristics, which enables discrimination from the other domains.

5.3. *RQ3: Which measures show the best performance to discriminate knowledge domains?*

To recall, with this question, we aim at finding the most essential (RDF) graph measures able to discriminate knowledge domains efficiently and to measure individual measure performance. We used the approach of setting up two classification tasks with Random Forest classifiers, each tuned by hyperparameter grid-search. The first task (1) is a multiclass classification problem, the second task (2) a two-class, one-vs-rest, binary version of the first. We removed three categories and the corresponding datasets from the initially available nine knowledge domains, due to too little datasets in these categories (<= 6, cf. Table 2). The remaining data was subject to standardization with robust-scaling since earlier, we found that most features have outliers.

Overall measure importance

Figure 6 shows the results of classification task (1). The colors encode graph measures in light-blue and RDF graph measures in blue. The x-axis shows all measures $m \in M'$. The y-axis shows the mean importance score obtained from 300 estimators' feature importance calculation, in descending order. It can be interpreted as a percentage value of the extent to which a particular feature contributes to decrease the weighted impurity in the decision tree.

While the ranking shows a steadily decreasing order, the overall scores are rather low. The first thirteen measures can be considered to have some impact. From the fourteenth value on, there is hardly a change, and the impact score is low.

Among the top ten measures of the highest score are three graph measures (`pseudo_diameter`, `coefficient_variation_out_degree`, and `distinct_classes`) and seven RDF graph measures. Overall, measures employing the out-degree are favored. `mean_predicate_list_degree`, describing the mean number of repeated predicate used to describe subjects, has the highest score; `m`, describing the number of edges, the lowest.

Per-category measure importance

Figure 7 shows the results of classification task (2), where one can get a picture on measure performance in each of the categories. It shows per knowledge domain the top seven measures with the highest scores from the one-vs-rest classifier. Like in Figure 6, the y-axis shows the degree of contribution to decrease impurity in the decision tree.

At first glance, we can see that the set of measures considered most important varies much across knowledge domains and that individual scores are higher than in classification task (1). Overall, there are thirteen distinct measures considered here (after measure selection, the initial set of measures in M' was twenty-nine). Among these, six measures are employing the *max*, three measures employing the

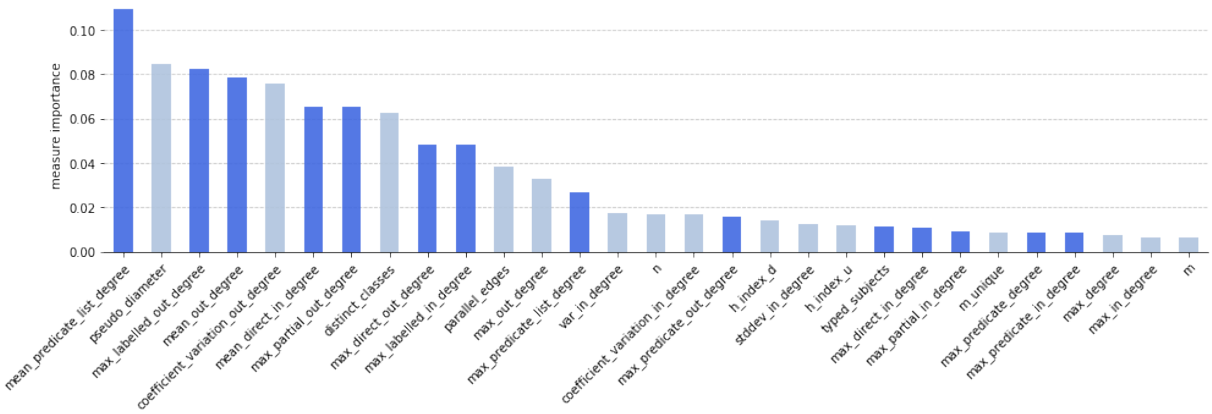

Fig. 6. **Overall measure importance** while discriminating datasets (classification task (1)). Shown are mean values for all non-redundant measures $m \in M'$. Colors encode graph measures (in light-blue) and RDF graph measures (in blue).

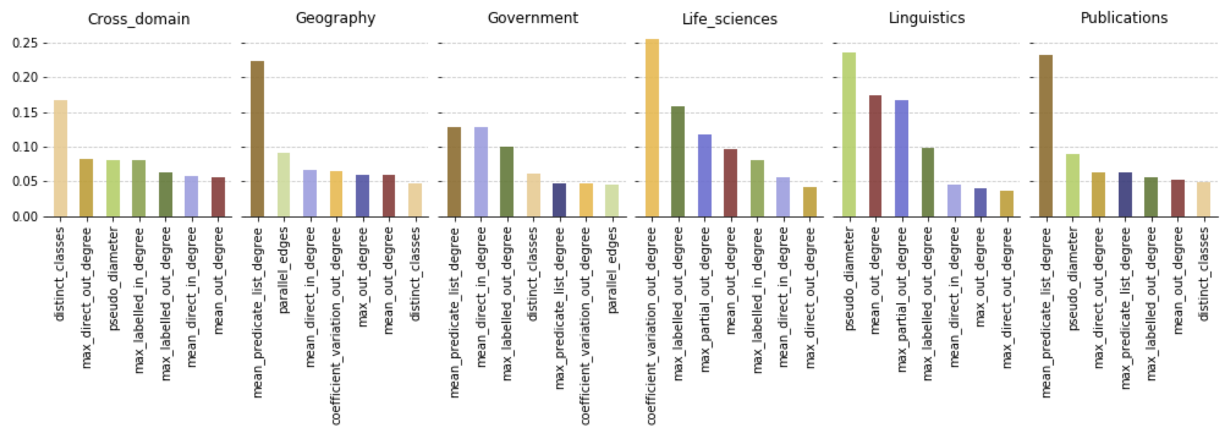

Fig. 7. **Per-category measure importance** while discriminating datasets (classification task (2)). Measures are encoded by color throughout all knowledge domains.

mean, and four measures employing an other absolute value. To complete this overall observation: out of the thirteen distinct measures, six employ outgoing edges, i.e., RDF predicates of subjects; two employ incoming edges of objects.

`max_labelled_out_degree`, `mean_direct_in_degree` and `mean_out_degree` are present in five out of six knowledge domains, although each with different scores and ranking. `distinct_classes` and `max_direct_out_degree` are present in four domains. Collecting the top two and top three measures of each knowledge domain results in having ten and eleven distinct measures, respectively. No measure is present in exclusively one category. Hence, there seems to be no measure with particular importance in a specific category. However, `distinct_classes`, `coefficient_variation_out_degree`, and `pseudo_diameter`, have highest scores in *Cross Domain*, *Life Sciences*, and *Linguistics*, respectively. `mean_predicate_list_degree` is even scored highest in three domains: *Geography*, *Government*, and *Publications*. By far, `mean_predicate_list_degree`, `coefficient_variation_out_degree` have the highest scores in *Geography*, *Life Sciences*, and *Publications*, respectively. These measures can be considered most important in the corresponding categories. Their values have distinctive characteristics, which enable classifiers to discriminate datasets according to these categories. Looking closer, the results of this binary relevance task here aligns well with the single performance analysis from above: the top ten measures from Figure 6 are the ones which are most likely to be found in the corresponding categories in Figure 7.

Summary of results

- To discriminate knowledge domains from each other, classifiers favor RDF graph measures over topological graph measures.
- Measures employing a max-value are favored over mean- and absolute values, like `distinct_classes`.
- Measures employing the out-degree are considered more important than measures employing the in-degree.
- To discriminate datasets from another, each knowledge domain considers a different set of measures as meaningful.

6. Discussion

We would like to address two major aspects exposed by the conducted experiments, namely (i) structural differences about RDF graphs from the viewpoint of graph measures, and (ii) the assessment of graph measure efficiency. The section closes up with limitations of this study.

6.1. Structural Characteristics of real-world RDF Datasets

The following discussion is based on the results of measure correlation coefficients (cf. Figure 1) and measure performance scores (cf. Figure 6 and 7).

General observations

By identifying effective graph features describing and discriminating RDF datasets and applying such features to LOD datasets, we gained an understanding of the topological differences of real-world datasets within distinct categories. The topology of RDF graphs (knowledge graphs more generally speaking) is distinct from other graph datasets, such as social graphs, due to the prevalence of hierarchical relations, that is, relations within the TBox (e.g. rdfs:subClassOf) or between ABox and TBox (e.g. rdf:type). This complements traversal relations and, by this means, imposes special characteristics that lead to generally higher connectivity, shorter paths, and the existence of vertex-“hubs” with high attractiveness from other vertices.

This is very well reflected in the graph measures. For example, measures like the number of edges, the maximum degree, and the maximum in-degree perfectly correlate with each other (cf. Section 5.1). Looking closer at the values for those measures reveals that 83% of the RDF graphs have vertices with a maximum in-degree being exactly equal to the maximum degree (in 94% of the cases, it is even almost equal). In most graphs, vertices representing the type (vertices with an “RDF type”-edge incident) are the ones with the highest in-degree. Such behavior of modeling, which is typical for RDF graphs and generally accepted as best practice in the RDF community, involves high connectivity of the graph’s topology. More references to the schema enhance this effect. In turn, more profound is the loss of connectivity as soon as the graph misses/loses references to the schema.

As more vertices and edges adhere to the graphs, the more heterogeneous and unstable the connectivity becomes. As a consequence, the overall density shrinks (cf. negative correlation of `m`, `max_degree` with `fill`) and the tendency of the topology to generate large sub-graphs having the shape of a “star” increase. Due to this and the aforementioned topological characteristic, measures employing the in-degree (some descriptive statistical measures, predicate (list-) degree measures, typed subjects, etc.) show a high correlation among each other. A stable value with growing size and volume of the graph would result in a homogeneous distribution, leading to a more stable and equally distributed connectivity of vertices among each other. The two mentioned examples can be considered being particularly RDF graph specific phenomena, which can be measured with the provided graph measures.

Observations within distinct categories

Vocabulary usage has a significant impact on the graph’s topology since schema and cardinality definitions are directly reflected in the graphs as options/restrictions to append vertices and edges. Thus, some measures are considered having a particular impact in individual categories, as shown in Figure 7. *Cross Domain*, for instance, has a diverse and irregular vocabulary usage, which implies a large number of mixed and heterogeneous datasets, with (larger) co-occurrence of schema references and type-statements (`distinct_classes`). *Geography* and *Publications* report on a regular usage of vocabularies. The recurrence of a fixed set of predicates (mean_predicate_list_degree) is the main distinguishable feature of these categories. *Geography* additionally reports on a proportionally high ratio of parallel edges of its datasets. Inherently, datasets in *Linguistics* stand out with a significantly larger path length of traversal relations (pseudo_diameter[6]). The modeling strategy there seems fairly concise, resulting in a low average number of types and outgoing predicates/edges per subject, which is reflected by the measures mean_out_degree and max_partial_out_degree.

In general, measure importance per category has a dependency to the way how publishers, data extraction tools, and researchers describe data. For example, according to the naming pattern datasets in *Linguistics* are clustered into three groups: *universal-dependencies-treebank-...* (63 datasets), *apertium-rdf-...* (22 datasets), and other (37 datasets). Other examples of clusters can be found in *Life Sciences* (*bio2rdf-...*, 26 datasets) and *Publications* (*rkb-explorer-...*, 32 datasets) categories. This implies similarities of vocabulary usage, which in turn is reflected in recurrences of particular patterns in the topological structure. On account of this fact, the prevalent measure impact is also influenced by the habits of people and tools populating datasets in the individual categories.

Therefore, category-specific topological characteristics should be reflected in samples, benchmarks, or synthetic data.

6.2. Efficient RDF Graph Measures

The initial set of fifty-four measures (M) was subject to correlation coefficient analysis and feature selection methods. The size of the set reduced to twenty-nine non-redundant measures after feature elimination (M'). This set was subject to an analysis of variability within and across knowledge domains. After this preliminary analysis, we employed a classifier to obtain feature impact scores to rank measure importance.

Both experiments in Section 5.3 evaluated the same distinct set of measures. Measures below the threshold of 0.02 were considered having a particularly low level of impact. From a mixed set of graph and RDF graph measures, we identified a final effective set of *thirteen* measures, that is distinct and meaningful.

Low variability

As mentioned earlier, datasets in the individual knowledge domains show similarities in their topological structure. Thus, the set of measures considered being efficient and meaningful varies across these categories (cf. Figure 7). According to the classifier, each of the thirteen measures provides some form of information gain and meaning. A somewhat naive intuition is that a measure with low variability is characteristic and may be a suited candidate in a particular category. The first experiment showed that measures with low variability (e.g., mean_out_degree, mean_direct_in_degree and pseudo_diameter) were preferred during category prediction and evaluated with higher impact scores (cf. Figure 5 and 6). The second experiment, focusing on individual categories, showed a different situation. Measures were considered characteristic and assessed with higher impact scores as their per-category variability (shown in Figure 4) was high. For example, mean_predicate_list_degree shows a high impact score in *Government* due to higher variability within and across categories (cf. Figure 4 and 5). Similar applies for other measures, like coefficient_variation_out_degree, max_partial_out_degree, and max_labelled_out_degree. *Cross Domain*, for instance, employs only measures of low variability (e.g., distinct_classes, max_direct_out_degree, etc.). Thus, in our classification tasks, the classifier tries to find the right balance between a low variability across categories and a somewhat characteristic variability as a topological feature.

Type of measures

Compared to other types of graphs, like social networks, RDF knowledge graph topologies adhere special characteristics, such as the pervasive reference to schema elements, with rdf:type statements being the most famous reference. This peculiarity influences the assessment about the meaningfulness of measures with regard to the discrimination of categories. For example, the classification task in Section 5.3 showed that RDF graph measures are preferred and obtained higher scores over other graph invariants, such as *h*-index (cf. Figure 6). Out of ten top-performing features in classification task (1), seven were RDF graph measures. Further, measures employing the in-degree are considered less effective, due to their heterogeneous ("unstable") value distributions. Hence, measures considering subjects and their out-degrees are considered more meaningful. Measures like the number of (parallel/unique) edges, maximal (in-) degree, maximum predicate (in-/out-) degree, and the number of typed subjects, are inherently high in variability within and across knowledge domains. Their heterogeneous character lets them be ineffective and not appropriate for dataset/category discrimination.

6.3. Limitations

There are some limitations of our experimental study that are worth to mention.

Size of the sample

The analysis of measure efficiency involved 280 datasets out of $1,163$ (end of 2017). While this number seems low regarding the theoretically available number of datasets, compared to other qualitative studies on datasets from the LOD Cloud, for instance [17–19], it sounds reasonable and of sufficient representativeness. Unfortunately, this is the current situation and, without additionally querying SPARQL-endpoints, the most that one can get from crawling the LOD Cloud.

Unbalanced domain classes

In order to tackle the class imbalance of our sample, we investigated class weighting and over- and undersampling techniques on the training sample passed to the classifier. Oversampling creates synthetic datasets (no duplicates) in each class up to the number of datasets of the largest class; undersampling down-sampled all classes to the size of the smallest class.

Feature importance methods are sensitive to the data structure and the distribution of feature values, and thus all methods showed different scores for the corresponding measures. What is interesting though, the set of measures considered important was similar to a great extent, in particular the most important measure per category (e.g., `mean_out_degree`, `mean_predicate_list_degree`, `pseudo_diameter`, and `max_labelled_out_degree`). Further, the model was trained following best practices for model tuning and cross-validation-based model selection. Hence, we assume that the obtained impact ratio of the classifier for each feature is reliable.

Limited set of features

If one actually wanted to perform category prediction [34, 35] or measure the structural similarity between RDF datasets [36], we could ask if the graph measures presented in this paper are appropriate and sufficient. As discussed earlier, vocabulary usage and the way how publishers, data extraction tools, and researchers describe data, has an impact on the graph's topology. Employing merely *ontological* information of the RDF dataset is, however, not sufficient to reach acceptable prediction accuracy [34]. Our classification experiment showed that, by employing *topological* measures, the prediction of categories for datasets is possible. Thus, knowledge domain-related, topological, and dataset features should complement one another. Aligning and integrating other tools and features for the extraction of metadata and vocabulary usage [4] would achieve improvements in prediction accuracy. Further, the integration of measures to somewhat distinguish hierarchical and traversal relations in the graphs, as this is a key characteristic for RDF data, would be beneficial.

Application and generalization of the findings to other (non-LOD) graphs

All of the measures in M and M' can be computed on RDF graphs and datasets outside of the LOD Cloud. Although metrics introduced by Fernández et al. [3] are considered for RDF graph characterization, of which in this paper only some could be implemented into our framework[3] and included in the study about measure efficiency, on closer inspection, most of them could also be applied to non-RDF graphs. `distinct_classes`, `typed_subjects`, and `ratio_of_typed_subjects` form exceptions, as they require edges explicitly labeled with `rdf:type`. To analyze non-RDF graphs, an essential requirement is to have some form of *consistent* labeling (literal or numeric) of the edges during graph initialization.

However, as RDF graphs are multigraphs, which may contain multiple edges between the same pair of source and target vertices, and whose use of (partly) very specialized vocabularies exposes special characteristics to the graph's topology, the results are unlikely to be applicable to non-RDF graphs and categories outside the LOD Cloud. However, this should be an interesting research question.

7. Conclusion and Future Work

We have created a framework with which one may efficiently compute topological graph measures for an arbitrary number of RDF datasets [2]. The main objective of this paper is to assess measure effectiveness and performance of fifty-four graph and RDF graph measures for RDF datasets. This is accomplished by means of statistical tests, such as the analysis of correlation coefficients, results of feature selection, variability, and a supervised classification task, in order to assess a measure's efficiency in terms of its capacity to discriminate dataset knowledge domains. For this purpose, a sample of 280 RDF datasets from nine knowledge domains was acquired from the LOD Cloud late 2017. All 280 datasets, instantiated graph objects, and values for fifty-four measures per graph are available for download on our website[13]. Please note that, despite following best practices for model tuning and cross-validation-based model selection, the primary aim was not to find the best classification model but to provide an understanding of feature performance, i.e., the importance of distinct graph measures in this particular task.

From a mixed set of initially fifty-four graph and RDF graph measures, the final set of *thirteen* measures is actually effective, distinct, and meaningful, in order to describe RDF graphs. The majority of the measures are RDF graph-based, according to the definition in [3], and preferably employs the out-degree and out-going edges of subjects to some extend. To discriminate categories, the following measures have the most significant impact: the average number of repeated predicate lists (`mean_predicate_list_degree`), the diameter of the graph (`pseudo_diameter`[6]), the maximum number of predicates with which a subject is related (`max_labelled_out_degree`), and the mean out-degree of the vertices (`mean_out_degree`).

The prevalent structure of topology is shaped by means of two mutually influencing aspects: (1) fundamental characteristics that adhere to RDF knowledge graph topologies in particular, and (2) the compliance to a standardized vocabulary. The distinctness of a measure's impact in the individual knowledge domains implies that there are fundamental differences in the shape of topologies. An RDF dataset that is re-using a popular vocabulary will likely show characteristics that can be found in other RDF graphs. The more diverse the use of vocabularies in a dataset is, the more variety and irregularity will be found in common structural patterns of the topology. Therefore, datasets using proprietary vocabularies will differ in their structure. Hence, a group of RDF graphs with similar characteristics causes knowledge domain-dependent feature performance and impact.

Apart from the classification experiments, we also gained some understanding of the general ability to predict category labels for RDF datasets, by relying on topological measures of the graphs exclusively. The positively surprising accuracy is comparable with other approaches and experiments, such as [34] and [35]. We came to the conclusion that this is on account of the usage of standardised and established vocabularies in the knowledge domains itself. This can be considered as being a qualitative aspect of a particular knowledge domain.

Implications

We are confident that related work in the fields of *synthetic dataset generation*, *sampling methods*, and frameworks for *quality evaluation*, e.g., can benefit from considering efficient topological (RDF) graph measures and category-specific assessments of the RDF graph's topology.

- A primary goal of synthetic dataset generators is to emulate datasets and to be as close as possible to a real-world setting. Thus, topological characteristics exhibited by a particular knowledge domain are of high value. Beyond parameters like the dataset size, which is typically interpreted as the number of triples, synthetic dataset generators might employ meaningful and disregard non-efficient (RDF) graph measures, in order to target the domain of test-data generation more appropriately.
- Sampling methods aim at finding a most representative sample from an original dataset. Apart from considering qualitative aspects, like classes, properties, instances, and used vocabularies, also topological aspects of the original RDF graph should be considered. Our framework and the proposed (RDF) graph-based measures could help to evaluate the quality of a graph sample.
- Having topological measures as another group of features is beneficial for solutions that evaluate and ensure the quality of Linked Open Data, such as RDF dataset profile generators. Concerning efficient measures, each category (LOD Cloud domain class) might have its own

[13]https://data.gesis.org/lodcc/2017-08

understanding of quality, such as a large diameter for datasets in *Linguistics*, a lower average degree for datasets in the *Life Sciences*, etc. Outliers and striking values for some measures could be indicators for erroneous data or ways of modeling or using a vocabulary that is not compliant with the knowledge domain of interest.

Future work

Part of our future work is to align graph features with features extracted by established RDF profiling tools. This widens the field of potential research and applications involving graph-based measures. For instance, we plan to improve the prediction of appropriate category labels for datasets by including features at instance- and schema-level of an RDF dataset. Moreover, this enables research in the direction of quality assurance and dataset search. We further plan to include more datasets from sources, e.g., SPARQL endpoints and non-LOD Cloud datasets. The evaluation of measures will be extended towards non-RDF graphs, with the aim to compare measure impact between these two types of graphs.

In terms of infrastructure, our portal is going to be updated with an upload functionality. A website visitor may then upload or provide the URL of an RDF dataset to let our framework analyze the corresponding RDF graph. By this means, we hope to collect more datasets and statistics.

In order to facilitate the access, usage, and querying of the results, we consider to represent all measures for all RDF graphs as an RDF dataset itself and import it into a publicly available SPARQL-endpoint. The RDF Data Cube Vocabulary [37] is considered for this.

References

[1] M.E.J. Newman, *Networks: An Introduction*, Oxford University Press, 2010. ISBN 9780198805090.

[2] M. Zloch, M. Acosta, D. Hienert, S. Dietze and S. Conrad, A Software Framework and Datasets for the Analysis of Graph Measures on RDF Graphs, in: *ESWC*, P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A.J.G. Gray, V. López, A. Haller and K. Hammar, eds, Lecture Notes in Computer Science, Vol. 11503, Springer, 2019, pp. 523–539. ISBN 978-3-030-21348-0.

[3] J.D. Fernández, M.A. Martínez-Prieto, P. de la Fuente Redondo and C. Gutiérrez, Characterising RDF data sets, *Journal of Information Science* **44**(2) (2018), 203–229.

[4] M. Ben Ellefi, Z. Bellahsene, B. John, E. Demidova, S. Dietze, J. Szymanski and K. Todorov, RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications, *Semantic Web journal* **9**(5) (2018), 677–705.

[5] M. Zneika, D. Vodislav and D. Kotzinos, Quality Metrics for RDF Graph Summarization, *Semantic Web journal* **10**(3) (2019), 555–584.

[6] M.P. Consens, V. Fionda, S. Khatchadourian and G. Pirrò, S+EPPs: Construct and Explore Bisimulation Summaries, plus Optimize Navigational Queries; all on Existing SPARQL Systems, *Proceedings of the VLDB Endowment (PVLDB)* **8**(12) (2015), 2028–2031.

[7] S. Campinas, T.E. Perry, D. Ceccarelli, R. Delbru and G. Tummarello, Introducing RDF Graph Summary With Application to Assisted SPARQL Formulation, in: *Proceedings of the 23rd International Workshop on Database and Expert Systems Applications*, IEEE Computer Society, Los Alamitos, California 2012, 2012, 11th International Workshop on Web Semantics.

[8] M. Schmachtenberg, C. Bizer and H. Paulheim, Adoption of the Linked Data Best Practices in Different Topical Domains, in: *ISWC - International Semantic Web Conference*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C.A. Knoblock, D. Vrandecic, P.T. Groth, N.F. Noy, K. Janowicz and C.A. Goble, eds, Lecture Notes in Computer Science, Vol. 8796, Springer, 2014, pp. 245–260. ISBN 978-3-319-11963-2.

[9] J. Demter, S. Auer, M. Martin and J. Lehmann, LODStats - An Extensible Framework for High-performance Dataset Analytics, in: *International Conference on Knowledge Engineering and Knowledge Management, EKAW*, A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles and N. Hernandez, eds, Lecture Notes in Computer Science, Vol. 7603, Springer, 2012, pp. 353–362. ISBN 978-3-642-33875-5.

[10] G. Sejdiu, I. Ermilov, J. Lehmann and M.N. Mami, DistLODStats: Distributed Computation of RDF Dataset Statistics, in: *ISWC - International Semantic Web Conference*, D. Vrandecic, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee and E. Simperl, eds, Lecture Notes in Computer Science, Vol. 11137, Springer, 2018, pp. 206–222. ISBN 978-3-030-00668-6.

[11] S. Khatchadourian and M.P. Consens, ExpLOD: Summary-Based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud, in: *The Semantic Web: Research and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 272–287.

[12] B. Fetahu, S. Dietze, B.P. Nunes, M.A. Casanova, D. Taibi and W. Nejdl, A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles, in: *ESWC*, V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab and A. Tordai, eds, Lecture Notes in Computer Science, Vol. 8465, Springer, 2014, pp. 519–534. ISBN 978-3-319-07442-9.

[13] C. Böhm, F. Naumann, Z. Abedjan, D. Fenz, T. Grütze, D. Hefenbrock, M. Pohl and D. Sonnabend, Profiling Linked Open Data with ProLOD, in: *ICDE Workshops*, IEEE Computer Society, 2010, pp. 175–178. ISBN 978-1-4244-6522-4.

[14] Z. Abedjan, T. Grütze, A. Jentzsch and F. Naumann, Profiling and mining RDF data with ProLOD++, in: *IEEE International Conference on Data Engineering, ICDE*, I.F. Cruz, E. Ferrari, Y. Tao, E. Bertino and G. Trajcevski, eds, IEEE Computer Society, 2014, pp. 1198–1201. ISBN 978-1-4799-3480-5.

[15] N. Mihindukulasooriya, M. Poveda-Villalón, R. García-Castro and A. Gómez-Pérez, Loupe - An Online Tool for Inspecting Datasets in the Linked Data Cloud, in: *ISWC - International Semantic Web Conference, (Posters & Demos)*, S. Villata, J.Z. Pan and M. Dragoni, eds, CEUR Workshop Proceedings, Vol. 1486, CEUR-WS.org, 2015.

[16] A. Hogan, A. Harth, A. Passant, S. Decker and A. Polleres, Weaving the Pedantic Web, in: *LDOW Workshop*, C. Bizer, T. Heath, T. Berners-Lee and M. Hausenblas, eds, CEUR Workshop Proceedings, Vol. 628, CEUR-WS.org, 2010.

[17] J. Debattista, C. Lange, S. Auer and D. Cortis, Evaluating the quality of the LOD Cloud: An empirical investigation, *Semantic Web journal* **9**(6) (2018), 859–901.

[18] J. Debattista, J. Attard, R. Brennan and D. O'Sullivan, Is the LOD Cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud, in: *WWW (Companion Volume)*, S. Amer-Yahia, M. Mahdian, A. Goel, G.-J. Houben, K. Lerman, J.J. McAuley, R. Baeza-Yates and L. Zia, eds, ACM Digital Library, 2019, pp. 850–858. ISBN 978-1-4503-6675-5.

[19] A. Haller, J.D. Fernández, M.R. Kamdar and A. Polleres, What are Links in Linked Open Data? A Characterization and Evaluation of Links between Knowledge Graphs on the Web, *Working Papers on Information Systems, Information Business and Operations* (2019).

[20] L. Ding and T. Finin, Characterizing the Semantic Web on the Web, in: *ISWC - International Semantic Web Conference*, I.F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold and L. Aroyo, eds, Lecture Notes in Computer Science, Vol. 4273, Springer, 2006, pp. 242–257. ISBN 3-540-49029-9.

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

[22] W. Hu, H. Qiu and M. Dumontier, Link Analysis of Life Science Linked Data, in: *ISWC - International Semantic Web Conference*, M. Arenas, Ó. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P.T. Groth, M. Dumontier, J. Heflin, K. Thirunarayan and S. Staab, eds, Lecture Notes in Computer Science, Vol. 9367, Springer, 2015, pp. 446–462.

[23] D.J. Watts, Networks, Dynamics, and the Small-World Phenomenon, *American Journal of sociology* **105**(2) (1999), 493–527.

[24] R. Gil and R. García, Measuring the Semantic Web, in: *Advances in Metadata Research. Proceedings of MTSR 2005*, Vol. 1, S. Sanchez-Alonso, ed., Rinton Press, Princeton, NJ, USA, 2006, pp. 72–77. ISBN 1-58949-053-3.

[25] D. Bachlechner and T. Strang, Is the Semantic Web a small world?, in: *ITA - International Conference on Internet Technologies and Applications (2)*, 2007, pp. 413–422. https://elib.dlr.de/47899/.

[26] A. Flores, M.-E. Vidal and G. Palma, Graphium Chrysalis: Exploiting Graph Database Engines to Analyze RDF Graphs, in: *ESWC (Satellite Events)*, V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis and A. Tordai, eds, Lecture Notes in Computer Science, Vol. 8798, Springer, 2014, pp. 326–331. ISBN 978-3-319-11954-0.

[27] C. Gutiérrez, C.A. Hurtado, A.O. Mendelzon and J. Pérez, Foundations of Semantic Web databases, *Journal of Computer and System Sciences.* **77**(3) (2011), 520–541.

[28] R. Cyganiak, D. Wood and M. Lanthaler (eds), RDF 1.1 Concepts and Abstract Syntax, 2014. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[29] S. Qiao and Z.M. Özsoyoglu, RBench: Application-Specific RDF Benchmarking, in *SIGMOD*, ACM Digital Library, 2015, pp. 1825–1838. ISBN 978-1-4503-2758-9.

[30] J.E. Hirsch, An index to quantify an individual's scientific research output, *Proc. National Academy of Sciences of the United States of America* **102**(46) (2005).

[31] L.C. Freeman, Centrality in social networks: Conceptual clarification, *Social Networks* **1**(3) (1979), 215–239.

[32] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical Report, Stanford InfoLab, 1999.

[33] J. Alstott, E. Bullmore and D. Plenz, powerlaw: a Python package for analysis of heavy-tailed distributions, *PloS one* **9**(1) (2014), e85777.

[34] A. Abele, Linked Data Profiling: Identifying the Domain of Datasets Based on Data Content and Metadata, in: *WWW (Companion Volume)*, J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks and B.Y. Zhao, eds, ACM Digital Library, 2016, pp. 287–291. ISBN 978-1-4503-4144-8.

[35] S. Lalithsena, P. Hitzler, A.P. Sheth and P. Jain, Automatic Domain Identification for Linked Open Data, in: *Web Intelligence*, IEEE Computer Society, 2013, pp. 205–212. ISBN 978-1-4799-2902-3.

[36] P. Maillot and C. Bobed, Measuring structural similarity between RDF graphs, in: *SAC*, H.M. Haddad, R.L. Wainwright and R. Chbeir, eds, ACM Digital Library, 2018, pp. 1960–1967.

[37] R. Cyganiak and D. Reynolds (eds), The RDF Data Cube Vocabulary, 2014. https://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/.

[38] L.-C. Zhang and M. Patone, Graph sampling, *METRON* **75**(3) (2017), 277–299. https://doi.org/10.1007/s40300-017-0126-y.

[39] G. Aluç, O. Hartig, M.T. Özsu and K. Daudjee, Diversified Stress Testing of RDF Data Management Systems, in: *ISWC - International Semantic Web Conference*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C.A. Knoblock, D. Vrandecic, P.T. Groth, N.F. Noy, K. Janowicz and C.A. Goble, eds, Lecture Notes in Computer Science, Springer, 2014, pp. 197–212. ISBN 978-3-319-11963-2.

[40] A. Assaf, R. Troncy and A. Senart, Roomba: An Extensible Framework to Validate and Build Dataset Profiles, in: *ESWC (Satellite Events)*, Lecture Notes in Computer Science, Vol. 9341, Springer, 2015, pp. 325–339.

[41] C. Bizer and A. Schultz, The Berlin SPARQL Benchmark, *International Journal on Semantic Web and Information Systems (IJSWIS)* (2009), 1–24.

[42] M.A. Bornea, J. Dolby, A. Kementsietsidis, K. Srinivas, P. Dantressangle, O. Udrea and B. Bhattacharjee, Building an Efficient RDF Store over a Relational Database, in: *SIGMOD International Conference on Management of Data*, SIGMOD, Association for Computing Machinery, 2013, pp. 121–132–. ISBN 9781450320375. https://doi.org/10.1145/2463676.2463718.

[43] S. Duan, A. Kementsietsidis, K. Srinivas and O. Udrea, Apples and Oranges: A Comparison of RDF Benchmarks and real RDF Datasets, in *SIGMOD*, ACM Digital Library, 2011, pp. 145–156.

[44] O. Erling, Virtuoso, a Hybrid RDBMS/Graph Column Store, *IEEE Database Engineering Bulletin* **35**(1) (2012), 3–8.

[45] Y. Guo, Z. Pan and J. Heflin, LUBM: A Benchmark for OWL Knowledge Base Systems, *Journal of Web Semantics* **3**(2) (2005), 158–182.

[46] J. Hayes and C. Gutiérrez, Bipartite Graphs as Intermediate Model for RDF, in: *ISWC - International Semantic Web Conference*, S.A. McIlraith, D. Plexousakis and F. van Harmelen, eds, Lecture Notes in Computer Science, Vol. 3298, Springer, 2004, pp. 47–61. ISBN 3-540-23798-4.

[47] A. Jentzsch, C. Dullweber, P. Troiano and F. Naumann, Exploring Linked Data Graph Structures, in: *ISWC - International Semantic Web Conference (Posters & Demos)*, S. Villata, J.Z. Pan and M. Dragoni, eds, CEUR Workshop Proceedings, Vol. 1486, CEUR-WS.org, 2015.

[48] J. Leskovec and C. Faloutsos, Sampling from large graphs, in: *SIGKDD - Conference on Knowledge Discovery and Data Mining*, T. Eliassi-Rad, L.H. Ungar, M. Craven and D. Gunopulos, eds, ACM Digital Library, 2006, pp. 631–636. ISBN 1-59593-339-5.

[49] F. Manola and E. Miller (eds), RDF Primer, W3C Recommendation, 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210.

[50] M.E.J. Newman, The structure and function of complex networks, in: *SIAM Review*, 2003, pp. 167–256.

[51] M.-D. Pham and P.A. Boncz, Exploiting Emergent Schemas to Make RDF Systems More Efficient, in: *ISWC - International Semantic Web Conference*, P.T. Groth, E. Simperl, A.J.G. Gray, M. Sabou, M. Krötzsch, F. Lécué, F. Flöck and Y. Gil, eds, Lecture Notes in Computer Science, Vol. 9981, 2016, pp. 463–479. ISBN 978-3-319-46522-7.

[52] Y. Roussakis, I. Chrysakis, K. Stefanidis, G. Flouris and Y. Stavrakas, A Flexible Framework for Understanding the Dynamics of Evolving RDF Datasets, in: *ISWC - International Semantic Web Conference*, M. Arenas, Ó. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P.T. Groth, M. Dumontier, J. Heflin, K. Thirunarayan and S. Staab, eds, Lecture Notes in Computer Science, Vol. 9366, Springer, 2015, pp. 495–512. ISBN 978-3-319-25006-9.

[53] M. Saleem, Q. Mehmood and A.-C.N. Ngomo, FEASIBLE: A Feature-Based SPARQL Benchmark Generation Framework, in: *ISWC - International Semantic Web Conference*, M. Arenas, Ó. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P.T. Groth, M. Dumontier, J. Heflin, K. Thirunarayan and S. Staab, eds, Lecture Notes in Computer Science, Vol. 9366, Springer, 2015, pp. 52–69. ISBN 978-3-319-25006-9.

[54] M. Schmidt, T. Hornung, G. Lausen and C. Pinkel, SP2Bench: A SPARQL Performance Benchmark, *Computing Research Repository (CoRR),* **abs/0806.4627** (2008). http://arxiv.org/abs/0806.4627.

[55] M.P.H. Stumpf, C. Wiuf and R.M. May, Subnets of scale-free networks are not scale-free: Sampling properties of networks, *Proceedings of the National Academy of Sciences of the United States of America* **102**(12) (2005), 4221–4224.

[56] Y.C. Tay, Data Generation for Application-Specific Benchmarking, *Proceedings of the VLDB Endowment (PVLDB), Challenges and Visions* **4**(12) (2011), 1470–1473.

[57] C. Wagner, P. Singer, F. Karimi, J. Pfeffer and M. Strohmaier, Sampling from Social Networks with Attributes, in: *WWW*, R. Barrett, R. Cummings, E. Agichtein and E. Gabrilovich, eds, ACM Digital Library, 2017, pp. 1181–1190. ISBN 978-1-4503-4913-0.

[58] K. Wilkinson, C. Sayers, H. Kuno and D. Reynolds, Efficient RDF Storage and Retrieval in Jena2, in: *Proceedings of the First International Conference on Semantic Web and Databases*, SWDB, CEUR-WS.org, 2003, pp. 120–139–.