Reachable Subwebs for Traversal-Based Query Execution

Olaf Hartig
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
ohartig@uwaterloo.ca

M. Tamer Özsu
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
tamer.ozsu@uwaterloo.ca

ABSTRACT

Traversal-based approaches to execute queries over *data on the Web* have recently been studied. These approaches make use of up-to-date data from initially unknown data sources and, thus, enable applications to tap the full potential of the Web. While existing work focuses primarily on implementation techniques, a principled analysis of subwebs that are reachable by such approaches is missing. Such an analysis may help to gain new insight into the problem of optimizing the response time of traversal-based query engines. Furthermore, a better understanding of characteristics of such subwebs may also inform approaches to benchmark these engines.

This paper provides such an analysis. In particular, we identify typical graph-based properties of query-specific reachable subwebs and quantify their diversity. Furthermore, we investigate whether vertex scoring methods (e.g., PageRank) are able to predict query-relevance of data sources when applied to such subwebs.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Miscellaneous

1. INTRODUCTION

In recent years, the publication of Linked Data on the World Wide Web (WWW) has gained significant momentum. Link traversal based query execution (LTBQE) approaches for live querying this emerging data space have received interest [5, 10, 11] since they do not depend on query processing functionality to be provided by data publishers; instead, they rely only on the lookup of URIs as a means of data access. The novelty of these approaches lies in integrating a traversal-based retrieval of data into the query execution process. Hence, these approaches do not assume a-priori a fixed set of potentially relevant data sources; instead, the traversal process discovers data and data sources on the fly.

While LTBQE approaches may answer queries based on data from initially unknown data sources, query results cannot guaranteed to be complete w.r.t. all Linked Data on the WWW [4]. As a consequence, LTBQE approaches typically support a reachability-based query semantics according to which the scope of any query is restricted to a well-defined subweb (that may differ for each query). For instance, under c_{Match}-semantics [4] such a subweb for a given query consists of all data that is reachable by traversing recursively all data links that match some pattern in the query.

Since LTBQE systems do not have a-priori information about the reachable subweb for any given query, to guarantee that a computed query result is complete, such a system has to fully explore the reachable subweb during query execution. These reachable subwebs may differ significantly among different queries. As a result, queries that are similar in a more traditional setting may cause dissimilar behavior when evaluated under a reachability-based query semantics over Linked Data on the WWW. For example, consider the following two SPARQL queries from the FedBench benchmark suite [13] (prefix declarations omitted).

LD_2:
```
SELECT * WHERE {
    ?proceedings swc:relatedToEvent
                 <http://data.semanticweb.org/conference/eswc/2010> .
    ?paper swc:isPartOf ?proceedings .    ?paper swrc:author ?p . }
```

LD'_{10}:
```
SELECT * WHERE {
    ?n dct:subject
       <http://dbpedia.org/resource/Category:Chancellors_of_Germany> .
    ?p2 owl:sameAs ?n .    ?p2 nyt:latest_use ?u . }
```

Both queries are structurally identical (i.e., both are path-shaped and have the same number and type of triple patterns). Thus, both queries appear to be similarly selective [14]. However, when we execute them (under c_{Match}-bag-semantics; cf. Section 2) using an LTBQE system that performs a breadth-first traversal strategy, we observe that executing LD_2 completely takes almost 5 times longer than executing LD'_{10} completely (109 min vs. 22 min), because the reachable subweb for LD_2 turns out to contain ca. 37 times more documents. On the other hand, we also notice that the query results for LD_2 and LD'_{10} are already complete after 3.7% and 63.6% of the overall query execution time, respectively (ca. 4 min vs. 14 min)!

These observations illustrate that the performance of any traversal-based execution of a given query depends significantly on the corresponding reachable subweb, and so does any attempt to optimize such an execution. Consequently, improving the state of the art in link traversal based query execution requires a detailed understanding of typical reachable subwebs and their properties.

To achieve such an understanding this paper presents a comprehensive analysis of various, query-specific reachable subwebs. In particular, this paper makes the following contributions:

1.) We study several graph-based properties of query-specific reachable subwebs and show that these subwebs may differ in multiple dimensions (i.e., not only in the number of documents covered).

2.) Based on these findings we introduce a quantitative approach to compare workloads of queries w.r.t. the diversity of their reachable subwebs. Such an approach is important for designing a realistic benchmark for testing LTBQE systems.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576947.

3.) Furthermore, we investigate whether methods for ranking graph vertices (such as PageRank) are suitable for predicting which of the documents in reachable subwebs actually contribute to the corresponding query result (in which case such a suitable method may be used for response time optimization in LTBQE systems).

Due to space limitations, this paper focuses on the concepts introduced for our analysis and *summarizes* the main findings. A more comprehensive technical report gives full account of our observations [6]. Furthermore, all digital artifacts related to our study (e.g., software, test data, etc.) are available online.[1]

The remainder of the paper is structured as follows: Section 2 introduces the formal foundations of our study. Sections 3 and 4 discuss graph-based properties of reachable subwebs and our approach to measure their diversity, respectively. Section 5 focuses on vertex-scoring methods and Section 6 concludes the paper.

2. PRELIMINARIES

This section introduces the formal foundations of our study. These foundations are based on a data model and a notion of SPARQL-based conjunctive queries under reachability-based semantics that we have formalized in our earlier work [4]. Due to space constraints, we omit most of the technical details of this formalization and define only the concepts used in the remainder of this paper.

Furthermore, we also assume familiarity with RDF [9] and the SPARQL query language [3]. We write $\mathcal{U}$, $\mathcal{B}$, $\mathcal{L}$, and $\mathcal{V}$ to denote the sets of all URIs, blank nodes, literals, and variables, respectively. Thus, a tuple from the set $\mathcal{T} = (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ is an *RDF triple* and a finite subset $B \subseteq (\mathcal{U} \cup \mathcal{V}) \times (\mathcal{U} \cup \mathcal{V}) \times (\mathcal{U} \cup \mathcal{L} \cup \mathcal{V})$ is a *basic graph pattern* (*BGP*), which is the basic building block of a SPARQL query. While SPARQL has further, more expressive features, existing work on LTBQE focuses on conjunctive queries expressed using BGPs [5, 10, 11]. Therefore, our study in this paper also focuses on BGPs. The standard SPARQL set semantics defines the query result of a BGP B over a set of RDF triples G as a set that we denote by $[\![B]\!]_G$ and that consists of partial mappings $\mu : \mathcal{V} \rightarrow (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$, which we refer to as *valuations*.

While the standard SPARQL semantics is suitable for querying a well-defined set of RDF triples (which might be stored in a DBMS), it is insufficient for querying Linked Data that is distributed over the WWW. Hence, to use BGPs as Linked Data queries in a well-defined manner, we need a query semantics that specifies the expected result of executing such a query over Linked Data on the WWW.

As a basis for defining such query semantics we have to introduce a **data model** that captures the idea of Linked Data formally: We define a *Web of Linked Data* as a tuple $W = (D, data, adoc)$ that consists of the following elements [4]: D is a set of symbols that we use to formally capture the concept of Web documents that can be obtained by looking up URIs in W. Hereafter, we call each $d \in D$ a Linked Data document, or *LD document* for short.

Mapping $data : D \rightarrow 2^{\mathcal{T}}$ associates each LD document with a finite set of RDF triples such that no blank node appears in $data(d)$ and in $data(d')$ of two distinct LD documents $d \neq d'$.

Finally, $adoc : \mathcal{U} \rightarrow D$ is a partial, surjective mapping which models the fact that a lookup of a URI $u \in \mathrm{dom}(adoc)$ in W results in the retrieval of LD document $adoc(u) = d \in D$. We may understand d as the authoritative source of data for URI u. Nonetheless, u may also be used in the data of other documents.

Then, using URI u in the data of LD document $d' \in D$ constitutes a *data link* to LD document $d = adoc(u)$. These data links form a graph structure that we call *link graph*. Formally, the link graph of a Web of Linked Data $W = (D, data, adoc)$ is a directed graph (D, E) whose vertices are all LD documents in W, and whose edges are all data links between these documents; i.e.,

$$E := \big\{ \big(d, adoc(u)\big) \in D \times D \,\big|\, t \in data(d) \text{ and } u \in \mathrm{uris}(t) \big\}.$$

Due to the openness and unbounded nature of the WWW, it is impossible to compute query results that are complete w.r.t. all Linked Data on the WWW [4]. Thus, to define queries that can be computed completely over any possible Web of Linked Data, we need a query semantics that restricts the scope of queries to well-defined "subwebs" of the queried Webs. However, a restriction to an a priori selected, fixed set of data sources significantly limits the possibilities for serendipitous discovery and, thus, does not allow Linked Data query execution systems to tap the full potential of the WWW.

Reachability-based query semantics avoid this limitation by using a notion of reachability to restrict the scope of a query. To specify such a notion of reachability formally, we have introduced the concept of a *reachability criterion* [4]. Given such a reachability criterion c, we have defined the reachable subweb of a queried Web of Linked Data in the context of a BGP B and a finite set of URIs $S \subseteq \mathcal{U}$ (which serve as a "seed"): Informally, the *(S, c, B)-reachable subweb* of a Web of Linked Data $W = (D, data, adoc)$ is a Web of Linked Data $W^* = (D^*, data^*, adoc^*)$ such that $D^* \subseteq D$ and any LD document $d \in D^*$ can be obtained using a seed URI $u \in S$—in which case we call d a *seed document*—or there exists a path in the link graph of W from a seed document to d such that each of the data links on that path "qualifies" according to reachability criterion c (mappings $data^*$ and $adoc^*$ depend on $data$ and $adoc$ in the obvious way [4]). An example of a reachability criterion is c_{Match} according to which a data link qualifies if that link corresponds to a triple pattern in the given BGP B [4]. Hereafter, we refer to each $d \in D^*$ as a *reachable document*.

We are now ready to define conjunctive Linked Data queries (CLD queries) that use BGPs under a reachability-based query semantics: The *CLD query* that uses a BGP B, a set of seed URIs S, and reachability criterion c, denoted by $\mathcal{Q}_c^{B,S}$, is a total function over the set of all Webs of Linked Data; for any such Web W, this function is defined by $\mathcal{Q}_c^{B,S}(W) := (\Omega, \rho)$ such that (Ω, ρ) is a multiset of valuations whose underlying set is $\Omega := [\![B]\!]_{\mathrm{AllData}(W^*)}$ with $W^* = (D^*, data^*, adoc^*)$ being the (S, c, B)-reachable subweb of W and $\mathrm{AllData}(W^*) = \bigcup_{d \in D^*} data(d)$; the corresponding function $\rho : \Omega \rightarrow \{1, 2, ...\}$ defines the cardinality of each valuation $\mu \in \Omega$ in the multiset as the number of distinct mappings $prv : \mu[B] \rightarrow D^*$ such that $t \in data\big(prv(t)\big)$ for all $t \in \mu[B]$.

We emphasize that the given definition of CLD queries introduces a family of (reachability-based) query semantics, each of which is based on a different reachability criterion c and, hereafter, referred to as *c-semantics*. Observe that all these query semantics are *bag semantics* (which is a divergence from our earlier work in which we define set semantics [4]). Hence, valuations may appear multiple times in a query result because any RDF triple used for constructing such a valuation may occur in the data of more than one (reachable) LD document. Bag semantics are more suitable for our study (and usually more easy to implement in systems).

Finally, we note that existing work on LTBQE techniques focuses on CLD queries under c_{Match}-semantics (or slight variations thereof) [5, 10, 11]. Therefore, our analysis in this paper also focuses c_{Match}-semantics. However, the concepts that we shall define for our analysis are generic and, thus, may be applied easily to analyses that focus on any other reachability-based semantics.

3. PROPERTIES OF LINK GRAPHS

We now study typical graph-based properties of reachable subwebs. For this study we are interested in some aspects of such

[1] http://squin.org/experiments/WWW2014/

Query (BGP)	#docs	#edges	e/v	#sc-comp	diameter	acyc.	#rlv-docs	%rlv-docs	res-size
LD_1	5167	13844	2.679	4	50	no	314	6.08 %	5662
LD_2	10175	28453	2.796	6	31	no	237	2.33 %	1480
LD_4	12896	43131	3.345	6	26	no	61	0.47 %	1600
LD_5	74	193	2.608	2	3	no	49	66.22 %	180
LD'_6	9655	13245	1.372	8140	7	no	66	0.68 %	1044
LD'_7	18	34	1.889	1	3	no	18	100.00 %	64
LD'_9	1710	10514	6.149	1	26	no	3	0.18 %	4
LD'_{10}	1554	1920	1.236	1457	6	no	7	0.45 %	12
wstdev:	57301.44	141501.07	14.52	18655.64	143.30	n/a	1086.12	362.94 %	15983.53
wstdev $P_{\mathbf{SQ3}}$:	1937.32	3755.55	8.78	673.78	11.19	n/a	28.11	30.95 %	46.04
rel. difference:	0.966	0.973	0.396	0.964	0.922	n/a	0.974	0.915	0.997

Table 1: Characteristics of query profile graphs (QPGs) for some of the FedBench Linked Data queries over the WWW.

subwebs that go beyond what is captured by the link graph of these subwebs. More precisely, in addition to information about how the reachable LD documents are interlinked with each other, we are interested in (i) the relevance and the (relative) importance of those LD documents for the corresponding query result and (ii) whether those LD documents are seed documents. Consequently, to capture all information relevant for our study, we introduce a more enhanced graph structure that extends the notion of the link graph of a reachable subweb as follows: Given a CLD query $\mathcal{Q}_c^{B,S}$, a Web of Linked Data $W = (D, data, adoc)$, and the (S, c, B)-reachable subweb of W, denoted by W^*, the *query profile graph* (*QPG*) of $\mathcal{Q}_c^{B,S}$ over W is a multirooted, vertex-weighted directed graph $p = (D^*, E, R, rcc)$ such that:

- (D^*, E) is the link graph of reachable subweb W^*;
- the set of root vertices $R \subseteq D^*$ are the *seed documents* (for $\mathcal{Q}_c^{B,S}$ in W); i.e., $R := \{adoc(u) \in D^* \,|\, u \in S\}$; and
- the vertex labeling function $rcc : D^* \to \{0, 1, 2, ...\}$ maps each LD document $d \in D^*$ to the number of valuations in $\mathcal{Q}_c^{B,S}(W)$ whose computation is based on an RDF triple in $data(d)$; i.e., $rcc(d) := \big|\{\mu \in \Omega \,|\, \mu[B] \cap data(d) \neq \varnothing\}\big|$, where Ω is the underlying set of $\mathcal{Q}_c^{B,S}(W) = (\Omega, \rho)$.

We call $rcc(d)$ of an LD document $d \in D^*$ the *result contribution counter* of d, and d is *relevant* (for $\mathcal{Q}_c^{B,S}$ over W) if $rcc(d) > 0$.

For our study we executed various CLD queries and used information recorded during these executions to construct QPGs. In the following, we first discuss our observations for QPGs obtained from executing queries of the FedBench benchmark [13] over "real" Linked Data on the WWW. Afterwards, we analyze QPGs of additional test queries over different, artificial Webs of Linked Data.

3.1 FedBench Queries

The FedBench benchmark suite proposes to test Linked Data query systems using a set of eleven BGPs, LD_1, ..., LD_{11} [13]. Hence, these BGPs are designed to be evaluated over Linked Data on the WWW (notably, FedBench does not specify any query semantics for such an evaluation). For our study, we use these BGPs under c_{Match}-semantics; that is, we have eleven CLD queries $\mathcal{Q}_{c_{\mathsf{Match}}}^{\mathsf{LD}_1,S_1}$, ..., $\mathcal{Q}_{c_{\mathsf{Match}}}^{\mathsf{LD}_{11},S_{11}}$; as seed URIs S_i these queries use all subject-position and object-position URIs mentioned in the corresponding BGP LD_i.

Preliminary tests with these queries revealed that some of the original FedBench BGPs (namely, LD_6 to LD_{10}) use outdated vocabularies. To fix this problem we slightly adjusted these BGPs (without changing the intent of the queries or their structural properties). The resulting BGPs, denoted by LD'_6 to LD'_{10}, and the other, original FedBench BGPs are given in our technical report [6].

Table 1 reports properties of the QPGs that we obtained by executing our FedBench-based CLD queries over the WWW. Before discussing these properties, we emphasize that we conducted this experiment from Nov. 11 to Nov. 18, 2013. During these days—in fact, during the whole time of our work on this paper—we have not been able to execute the queries that use LD_3, LD'_8, and LD_{11}, without observing a great number of URI lookups that time out nondeterministically (due to temporarily unresponsive Web servers).[2] As an unfortunate consequence, we have to exclude the three queries from our study (and, hence, they are missing from Table 1).

The types of properties of any QPG $p = (D^*, E, R, rcc)$ in Table 1 are the following (ignore the additional rows at the bottom of the table for the moment):

#docs: the number of vertices; i.e., $|D^*|$
#edges: the number of edges; i.e., $|E|$
e/v: the ratio of edges per vertex; i.e., $\frac{|E|}{|D^*|}$
#sc-comp: the number of strongly connected components
diameter: the length of longest shortest path between vertices
acyc.: represents whether the graph is acyclic
#rlv-docs: the number of relevant documents; i.e., $|D_{\mathsf{rlv}}|$ where $D_{\mathsf{rlv}} = \{d \in D^* | rcc(d) > 0\}$
%rlv-docs: the percentage of relevant documents; i.e., $\frac{|D_{\mathsf{rlv}}| \cdot 100\%}{|D^*|}$

In addition to these properties, Table 1 reports the size of the query results returned after executing *completely* the given FedBench-based CLD queries (column *res-size*); since query results are multisets (Ω, ρ), we measure their size as $\sum_{\mu \in \Omega} \rho(\mu)$.

The values in Table 1 illustrate that the (measured) properties differ significantly across the studied reachable subwebs (except for acyclicity). While these differences are not entirely unexpected for some properties, we have been somewhat surprised by the differences for #sc-comp and %rlv-docs. Note that the standard deviation (which is 36.29%) for the eight %rlv-docs values is 164.58% of their arithmetic mean (22.05%), and there are two *"outliers"* (LD_5 and LD'_7) that are not within one standard deviation from the mean; for the #sc-comp values, the standard deviation (2665.09) is even 221.70% of the mean (1202.13) with one outlier (LD'_6).

Our measurements also explain why, in the experiment outlined in the introduction, the query result for LD_2 was complete already after an unexpectedly small 3.7% of the overall query execution time (in contrast to 63.6% recorded for LD'_{10}): For both queries all relevant LD documents are close to the seed document (at most two steps away in both cases). However, for LD_2, the comparably high diameter suggests that many of the irrelevant LD documents are farther away, which is not the case for LD'_{10} (the aforementioned Web page for this paper provides a visualization of the corresponding QPGs that verifies this explanation). Therefore, since we have used a breath-first traversal for this experiment, the relevant documents

[2] For this experiment we adhered to the usual politeness policy of requesting at most two URIs per second from each Web server [7].

ϕ_1	ϕ_2	#docs	#edges	e/v	#sc-comp	diameter	acyc.	#rlv-docs	%rlv-docs	res-size
0	0	19	18	0.947	19	0	yes	0	0.00 %	0
0	0.33	55	54	0.982	55	3	yes	0	0.00 %	0
0	0.66	25	24	0.960	25	3	yes	2	8.00 %	1
0	1	1	0	0.000	1	0	yes	0	0.00 %	0
0.33	0	162	215	1.327	109	6	no	3	1.85 %	4
0.33	0.33	160	216	1.350	103	6	no	0	0.00 %	0
0.33	0.66	119	189	1.588	48	6	no	3	2.52 %	4
0.33	1	64	122	1.906	5	6	no	5	7.81 %	6
0.66	0	295	496	1.681	93	6	no	3	1.02 %	4
0.66	0.33	209	363	1.737	54	6	no	5	2.39 %	8
0.66	0.66	231	428	1.853	34	6	no	4	1.73 %	6
0.66	1	118	232	1.966	4	6	no	3	2.54 %	2
1	irrel.	367	734	2.000	1	6	no	7	1.91 %	12
wstdev:		1937.32	3755.55	8.78	673.78	11.19	n/a	28.11	30.95 %	46.04

Table 2: Characteristics of query profile graphs for query SQ3 over test Webs that differ in their link structure.

have been among the first reachable documents to be discovered during the execution of LD_2, whereas, for LD'_{10}, the breath-first traversal discovered many irrelevant documents in the beginning.

3.2 Simulation-Based Experiments

The differences of the properties that we have measured for the FedBench QPGs raise the research question of whether these differences are an artifact of using different queries or whether such differences can also be observed when querying different Webs using the same query. While the FedBench-based CLD queries allow us to study QPGs over the particular Web of Linked Data that exists on the WWW (at the time of our experiments), to answer the given question we aim to study QPGs of test queries over multiple, differently structured Webs of Linked Data. To be able to meaningfully compare the QPGs of a test query over different Webs, we used a single base dataset to generate a set of synthetic test Webs.

We selected as base dataset the set of RDF triples that the data generator of the Berlin SPARQL Benchmark suite [1] produces when called with a scaling factor of 200. This set, hereafter denoted by G_{base}, consists of 75,150 RDF triples and describes 7,329 entities in a fictitious e-commerce scenario (including products, reviews, etc.). Each of these entities is identified by a single, unique URI. Let U_{base} denote the set consisting of these 7,329 URIs.

Each *test Web* generated from this base dataset is a Web of Linked Data $W_{\mathsf{test}} = (D, data, adoc)$ for which the following properties hold: (i) $|D| = 7,329$, (ii) $\mathrm{dom}(adoc) = U_{\mathsf{base}}$, (iii) $adoc$ is bijective, and (iv) $\bigcup_{d \in D} data(d) = G_{\mathsf{base}}$.

To distribute the RDF triples from the base dataset G_{base} over such a test Web, we partitioned G_{base} into 7,329 (potentially overlapping) subsets, each of which became the set $data(d)$ for a different LD document $d \in D$. Given that $G_{\mathsf{base}} \subseteq U_{\mathsf{base}} \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{L})$, we always placed any base dataset triple $(s,p,o) \in G_{\mathsf{base}}$ with $o \notin U_{\mathsf{base}}$ into the set $data\big(adoc(s)\big)$. For any of the other triples $(s,p,o) \in G_{\mathsf{base}} \cap U_{\mathsf{base}} \times \mathcal{U} \times U_{\mathsf{base}}$, we considered three options: placing (s,p,o) into both $data\big(adoc(s)\big)$ and $data\big(adoc(o)\big)$, into $data\big(adoc(s)\big)$ only, or into $data\big(adoc(o)\big)$ only. We applied a random approach for choosing among these three options: With probability ϕ_1 the triple was placed into both $data\big(adoc(s)\big)$ and $data\big(adoc(o)\big)$; otherwise, it was placed into either $data\big(adoc(s)\big)$ or $data\big(adoc(o)\big)$ (but not into both), where ϕ_2 is the probability for placing the triple into $data\big(adoc(s)\big)$. It is easy to see that the selected probabilities ϕ_1 and ϕ_2 impact the link structure of the resulting test Web. Therefore, we could obtain a diverse set of differently structured test Webs by systematically varying ϕ_1 and ϕ_2. In particular, we have used each of the twelve pairs $(\phi_1, \phi_2) \in \{0, 0.33, 0.66\} \times \{0, 0.33, 0.66, 1\}$ to generate twelve test Webs $W_{\mathsf{test}}^{(0,0)}, \dots, W_{\mathsf{test}}^{(0.66,1)}$, and we complemented them with the test Web $W_{\mathsf{test}}^{(1)}$ that we generated using probability $\phi_1 = 1$ (in which case ϕ_2 is irrelevant)—giving us 13 test Webs in total.

To query these test Webs, we used six CLD queries under c_{Match}-semantics. These queries, denoted by SQ1 to SQ6, differ w.r.t. their structural properties (shape, size, etc.). Due to space limitations, the remainder of this section focuses on discussing QPGs of query SQ3. Our technical report describes all six queries and the properties of their QPGs over any of our test Webs [6].

We executed query SQ3 over any of the aforementioned 13 test Webs to collect information for constructing the corresponding 13 QPGs. Table 2 lists the properties of these QPGs (again, ignore the additional row at the bottom of the table for the moment).

First, we observe that for some properties the measured values differ significantly (similar to the FedBench case). However, some properties appear to be more regular (in particular, e/v and diameter). We also note that, in contrast to the FedBench QPGs, some QPGs of SQ3 are acyclic. This holds in particular for the test Webs with $\phi_1 = 0$. In these Webs, there is not a single RDF triple that establishes a bidirectional data link (a.k.a. "back links"), which naturally introduce cycles in the link graph of reachable subwebs.

Furthermore, we notice that for the test Webs generated with a greater ϕ_2 the number of reachable documents decreases. We explain this phenomenon as follows: The seed URI of query SQ3 appears in the object position of a triple pattern in the BGP of SQ3 and the subject of that triple pattern is a variable. Therefore, in the test Webs in which the corresponding seed document has been generated with a greater ϕ_2, that seed document contains more data links that satisfy reachability criterion c_{Match} and, thus, there exist more paths to reachable documents from such a seed document.

Our observations for the other test queries, SQ1, SQ2, SQ4, SQ5, and SQ6, are similar [6]. That is, the properties of reachable subwebs and their link graphs depend significantly on how the queried data is interlinked. However, the overall diversity of the QPGs for SQ3 (or any of the other five queries) does not seem to be as high as the diversity of the eight FedBench QPGs. In the following section we propose a quantitative approach that verifies this hypothesis.

4. DIVERSITY OF WORKLOADS

Let a finite set of pairs $(\mathcal{Q}, W)$, with $\mathcal{Q}$ being a CLD query and W being a Web of Linked Data, be a *workload*. Then, our observations in the previous section suggest that the QPGs for all pairs $(\mathcal{Q}, W)$ in some workload are more diverse than the QPGs for some other workload. This section proposes a quantitative approach for comparing workloads w.r.t. this diversity; we then apply this approach for particular workloads (such as those discussed before).

The main application of our approach is to assess the suitability of possible benchmarks for testing LTBQE systems. Apparently, a QPG and its properties influences how the corresponding query might be executed and what effect possible query optimizations have. Consequently, workloads that induce more diverse QPGs, are more suitable for benchmarking LTBQE systems.

QPGs may differ along multiple dimensions. Our approach focuses on eight dimensions that correspond to properties reported in Tables 1 and 2. We define a measure of diversity for a given set of QPGs that may be applied separately to any of these dimensions. Thereafter, we specify how two sets of QPGs can be compared by taking into account their relative diversity in all eight dimensions.

Let $M = \{$#docs, #edges, e/v, #sc-comp, diameter, #rlv-docs, %rlv-docs, res-size$\}$ be types of properties of QPGs as specified in Section 3.1. For each such property $m \in M$, let $m(p)$ denote the value that a given QPG p has for property m, and let $\mathrm{avg}^m(P)$ and $\mathrm{stdev}^m(P)$ be the arithmetic mean and the standard deviation of the m-values of a given set of QPGs P, respectively.

Then, we measure the diversity of P w.r.t. m by the *weighted standard deviation* $\mathrm{wstdev}^m(P) := \big(\omega_1^m(P) + \omega_2^m(P)\big) \cdot \mathrm{stdev}^m(P)$ where the weight is the sum of (i) the number of unique values $m(p)$ across all $p \in P$, i.e., $\omega_1^m(P) := \big|\{m(p) \,|\, p \in P\}\big|$, and (ii) the number of the "outlier" QPGs $p \in P$ whose value $m(p)$ is not within one standard deviation from the arithmetic mean, i.e., $\omega_2^m(P) := \big|\{p \in P \,|\, \mathrm{stdev}^m(P) < \mathrm{abs}(\mathrm{avg}^m(P) - m(p))\}\big|$.

For instance, the first additional row at the bottom of Tables 1 and 2 provides the weighted standard deviations for the set of QPGs listed in each of the tables, respectively. By comparing these values we note that, for every property $m \in M$, the set of FedBench QPGs is more diverse w.r.t. m than the QPGs of query SQ3 over the 13 test Webs (even if the set of FedBench QPGs contains five QPGs less). Hence, the overall diversity of the set of FedBench QPGs is also greater (that is, if we take into account all eight properties).

However, for some other pair of sets of QPGs, the first set may be more diverse w.r.t. some of the properties, whereas the second is more diverse w.r.t. other properties. Even in such a case we aim to identify the set of QPGs that has a greater overall diversity. To this end, we add up the relative differences of the respective weighted standard deviations. That is, given two sets of QPGs P_1 and P_2, we first compute the *relative difference* for any property $m \in M$:

$$\mathrm{rdiff}^m(P_1, P_2) := \begin{cases} 0 & \text{if } \mathrm{wstdev}^m(P_i) = 0 \text{ for all } i \in \{1, 2\}, \\ \frac{\mathrm{wstdev}^m(P_1) - \mathrm{wstdev}^m(P_2)}{\max\left(\mathrm{wstdev}^m(P_1), \mathrm{wstdev}^m(P_2)\right)} & \text{else.} \end{cases}$$

We now define the *diversity of* P_1 *relative to* P_2 as the sum of these differences; that is, $\mathrm{div}(P_1|P_2) := \sum_{m \in M} \mathrm{rdiff}^m(P_1, P_2)$.

For instance, if P_{FedB} denotes the set of FedBench QPGs (as listed in Table 1) and P_{SQ3} denotes the set of QPGs of query SQ3 over the 13 test Webs (in Table 2), the diversity of P_{FedB} relative to P_{SQ3} is $\mathrm{div}(P_{\text{FedB}}|P_{\text{SQ3}}) = 7.14$ (the corresponding relative differences $\mathrm{rdiff}^m(P_{\text{FedB}}, P_{\text{SQ3}})$ are given in the last row of Table 1).

We emphasize that, for any two sets of QPGs P_1 and P_2, any relative difference $\mathrm{rdiff}^m(P_1, P_2)$ (for all $m \in M$) is a rational number in the interval [-1,1].[3] As a consequence, $\mathrm{div}(P_1|P_2)$ is a rational number in [-8,8]. If $\mathrm{div}(P_1|P_2) > 0$ (resp. < 0), then P_1 is more (resp. less) diverse than P_2. If $\mathrm{div}(P_1|P_2) = 0$, then P_1 and P_2 are equally diverse. The latter may not only be the case if $\mathrm{wstdev}^m(P_1) = \mathrm{wstdev}^m(P_2)$ for all $m \in M$, but also if all eight relative differences $\mathrm{rdiff}^m(P_1, P_2)$ cancel out (when summed up).

Similar to comparing our (reduced) FedBench workload to the workload with query SQ3 (over our 13 test Webs), we compared the FedBench workload to workloads with our other five test queries. If $P_{\text{SQ}i}$ (for $i \in \{1, \dots, 6\}$) denotes the set of QPGs of test query SQi over the 13 test Web, the resulting relative diversities are:

$$\begin{aligned} &\mathrm{div}(P_{\text{FedB}}|P_{\text{SQ1}}) = 2.62, && \mathrm{div}(P_{\text{FedB}}|P_{\text{SQ2}}) = 5.72, \\ &\mathrm{div}(P_{\text{FedB}}|P_{\text{SQ3}}) = 7.14, && \mathrm{div}(P_{\text{FedB}}|P_{\text{SQ4}}) = 4.75, \\ &\mathrm{div}(P_{\text{FedB}}|P_{\text{SQ5}}) = 4.80, && \mathrm{div}(P_{\text{FedB}}|P_{\text{SQ6}}) = 6.80, \end{aligned}$$

Apparently, in all cases, the FedBench workload is more diverse.

Given this result, we are interested in how the FedBench workload compares to a workload that uses all six of our test queries over a single test Web. Thus, let $P_{(\phi_1,\phi_2)}$ denote the set that contains the six QPGs of any query SQ1, ... , SQ6 over the test Web $W_{\text{test}}^{(\phi_1,\phi_2)}$, respectively. We computed the following relative diversities:

$$\begin{aligned} \mathrm{div}(P_{\text{FedB}}|P_{(0,0)}) &= 6.13, & \mathrm{div}(P_{\text{FedB}}|P_{(0,0.33)}) &= 4.74, \\ \mathrm{div}(P_{\text{FedB}}|P_{(0,0.66)}) &= 4.67, & \mathrm{div}(P_{\text{FedB}}|P_{(0,1)}) &= 6.17, \\ \mathrm{div}(P_{\text{FedB}}|P_{(0.33,0)}) &= 4.07, & \mathrm{div}(P_{\text{FedB}}|P_{(0.33,0.33)}) &= 3.60, \\ \mathrm{div}(P_{\text{FedB}}|P_{(0.33,0.66)}) &= 3.81, & \mathrm{div}(P_{\text{FedB}}|P_{(0.33,1)}) &= 4.22, \\ \mathrm{div}(P_{\text{FedB}}|P_{(0.66,0)}) &= 3.17, & \mathrm{div}(P_{\text{FedB}}|P_{(0.66,0.33)}) &= 3.12, \\ \mathrm{div}(P_{\text{FedB}}|P_{(0.66,0.66)}) &= 3.15, & \mathrm{div}(P_{\text{FedB}}|P_{(0.66,1)}) &= 3.48, \\ \mathrm{div}(P_{\text{FedB}}|P_{(1)}) &= 2.64. \end{aligned}$$

Hence, even in these cases, the FedBench workload is more diverse.

5. VERTEX SCORING IN LINK GRAPHS

So far we have studied metrics that focus on a (link) graph as a whole. Now we turn to methods that assign some score to each vertex (e.g., PageRank). Hereafter, we refer to these methods as *vertex-scoring methods* or, simply *scoring methods*. For these methods we are interested in their suitability for predicting the (ir)relevance of reachable LD documents (which are the vertices in link graphs). As mentioned in the introduction, such a prediction may be used for response time optimizations in LTBQE systems. Therefore, this section first defines a quantitative approach for measuring whether a given vertex-scoring method is suitable; afterwards, we use this approach to evaluate several well-known vertex-scoring methods.

5.1 Measuring Suitability

Intuitively, a scoring method would be suitable for predicting the relevance of LD documents in a (query execution-specific) reachable subweb W^*, if there exists a correlation (or an anticorrelation) between the relevance (resp. the result contribution counter) of the LD documents and the scores that the method assigns to these LD documents in the link graph of W^*.

The typical approach to measure correlation is to use Pearson's correlation coefficient (PCC). However, in our scenario this approach is unsuitable because in many cases the percentage of reachable LD documents that are relevant is very small (as we have seen in Section 3). As a result, the few relevant LD documents appeared as outliers in a preliminary analysis during which we computed PCCs between the result contribution counter of reachable LD documents and some test scores. Therefore, we introduce an alternative approach to measure the suitability of vertex-scoring methods.

Given the QPG $p = (D^*, E, R, rcc)$ of a CLD query over a Web of Linked Data, and a scoring method sm, our approach consists of four steps: First, we use sm to compute the score of every non-seed document in the link graph (D^*, E). Let $score(d)$ denote this score for any non-seed document $d \in D^* \setminus R$ (we ignore the seed documents because LTBQE systems have already retrieved these seeds before they may begin prioritizing the lookup of discovered URIs). Hereafter, we use D_{ns} as shorthand for $D^* \setminus R$.

[3] For our use case, the relative difference is more suitable than the actual difference (i.e., $\mathrm{wstdev}^m(P_1) - \mathrm{wstdev}^m(P_2)$) because, e.g., an actual difference of 2 between values 1 and 3 is more significant than the same actual difference between values 101 and 103. The relative difference takes this significance into account.

Then, we normalize these scores to the interval $[0,1]$ (to make comparable our results for different scoring methods); that is, for each non-seed document $d \in D_{\mathsf{ns}}$, we compute a normalized score $nscore(d) := \frac{score(d) - min}{max - min}$ where $min = \min\big(\{score(d) \mid d \in D_{\mathsf{ns}}\}\big)$ and $max = \max\big(\{score(d) \mid d \in D_{\mathsf{ns}}\}\big)$.

The third step consists of computing the arithmetic mean of these normalized scores for the relevant non-seed documents and for all non-seed documents, respectively. Hence, we obtain

$$avg_{\mathsf{rel}} := \frac{\sum_{d \in D_{\mathsf{rel}}} nscore(d)}{|D_{\mathsf{rel}}|} \quad \text{and} \quad avg := \frac{\sum_{d \in D_{\mathsf{ns}}} nscore(d)}{|D_{\mathsf{ns}}|},$$

where $D_{\mathsf{rel}} = \{d \in D_{\mathsf{ns}} \mid rcc(d) > 0\}$.

We note that if avg_{rel} shows a clear tendency to be either notably high or low, then the scoring method sm may be suitable for predicting whether a reachable (non-seed) LD document $d \in D_{\mathsf{ns}}$ belongs to the set of relevant documents $D_{\mathsf{rel}} \subseteq D$. However, if there does not exist a significant difference between avg_{rel} and avg, the scoring method cannot be suitable (because, in this case, relevant and irrelevant documents are—on average—indistinguishable from each other w.r.t. the given score). Therefore, as the final step of our method, we compute the distance between both means, $dist := |avg_{\mathsf{rel}} - avg|$, and the difference of avg_{rel} to the center of interval $[0,1]$, $diff := avg_{\mathsf{rel}} - 0.5$.

If $dist < \alpha$ for a given threshold α, we say that the scoring method sm is $dist$-insignificant for QPG p. Similarly, if $|diff| < \beta$ for a given β, sm is $diff$-insignificant for QPG p. Then, based on the aforementioned reasoning, we conceive the scoring method sm as *unsuitable* for predicting the relevance of LD documents during an execution of query $\mathcal{Q}^{B,S}_{c_{\mathsf{Match}}}$ over W, if the sm is $dist$-insignificant or $diff$-insignificant for p (recall that the QPG p can be constructed only after executing $\mathcal{Q}^{B,S}_{c_{\mathsf{Match}}}$ over W).

By conducting such an analysis for a diverse set of QPGs, we may achieve an understanding of the general suitability (or unsuitability) of the scoring method sm for predicting the relevance of LD documents. In the following we describe such a study for several well-known scoring methods. For our study we use thresholds $\alpha = 0.25$ and $\beta = 0.1$ (note that $dist$ is a number in the interval [0,1] and $diff$ is in the interval [-0.5,0.5]).

5.2 Analyzing Well-Known Scoring Methods

Our study focuses on PageRank [12], HITS [8], k-step Markov [15], betweenness centrality [2], and the in-degree (i.e., the number incoming edges). We selected these methods because they present a mix of different types of vertex-scoring methods: PageRank and HITS are popular in the context of the WWW, k-step Markov is an example of measuring importance of vertices relative to some designated vertices, betweenness centrality is a global measure of vertex importance, and the in-degree has been used for prioritizing URI lookups in Ladwig and Tran's LTBQE approach [10].

Our analysis of these scoring methods shows that *none of them is suitable* [6], because, in most cases, they are $dist$-insignificant or $dist$ varies too much for different test Webs. For instance, the chart in Figure 1 illustrates $dist$ values that we measured for scoring method in-degree. Every cross in the chart represents the $dist$ measured for the corresponding test query, SQ1, ... SQ6, over one of our 13 test Webs. The dark blue dots represent the arithmetic mean of these measurements for each query and the error bars represent one standard deviation, respectively. Similar charts that illustrate the $diff$ values of the few $dist$-significant cases, show that these $diff$ values are also either insignificant or vary too much. Our results for the other scoring methods are very similar [6].

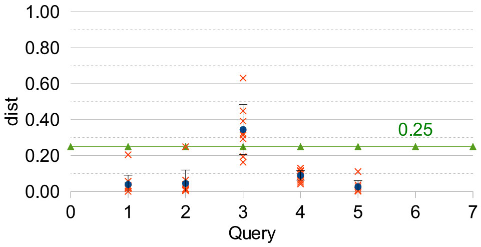

Figure 1: Chart that illustrates the values of $dist$ for in-degree.

We attribute the limited suitability of the studied vertex-scoring methods to the fact that none of these methods takes into account context-specific information about the LD documents for which they compute scores. Therefore, an interesting topic of future work is to develop new methods that use such information.

6. CONCLUSIONS

In this paper we have studied reachable subwebs of Linked Data queries under a reachability-based query semantics. We have shown that the subwebs for different queries may differ in multiple dimensions, and, even for the same query, reachable subwebs may differ significantly depending on how the queried Web is interlinked. Furthermore, we have proposed a quantitative approach to compare workloads of queries w.r.t. the diversity of their reachable subwebs and we have shown that well-known vertex-scoring methods are unsuitable for predicting query-relevance of data sources.

7. REFERENCES

[1] C. Bizer and A. Schultz. The Berlin SPARQL benchmark. *Semantic Web & Information Systems*, 5(2):1–24, 2009.

[2] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 1977.

[3] S. Harris, A. Seaborne, and E. Prud'hommeaux. SPARQL 1.1 query language. W3C Recommendation, Mar. 2013.

[4] O. Hartig. SPARQL for a Web of Linked Data: Semantics and computability. In *ESWC*, 2012.

[5] O. Hartig, C. Bizer, and J.-C. Freytag. Executing SPARQL queries over the Web of Linked Data. In *ISWC*, 2009.

[6] O. Hartig and M. T. Özsu. Reachable subwebs for traversal-based query execution. Technical Report CS-2014-02, University of Waterloo, Feb. 2014.

[7] A. Hogan, A. Harth, J. Umrich, S. Kinsella, A. Polleres, and S. Decker. Searching and browsing Linked Data with SWSE: the Semantic Web search engine. *Web Semantics*, 9(4), 2012.

[8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[9] G. Klyne and J. J. Carroll. Resource description framework (RDF): Concepts and abstract syntax. W3C Rec., Feb. 2004.

[10] G. Ladwig and D. T. Tran. Linked Data query processing strategies. In *ISWC*, 2010.

[11] D. P. Miranker, R. K. Depena, H. Jung, J. F. Sequeda, and C. Reyna. Diamond: A SPARQL query engine, for Linked Data based on the rete match. In *AImWD*, 2012.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, Nov. 1999.

[13] M. Schmidt, O. Görlitz, P. Haase, G. Ladwig, A. Schwarte, and T. Tran. FedBench: A benchmark suite for federated semantic data query processing. In *ISWC*, 2011.

[14] P. Tsialiamanis, L. Sidirourgos, I. Fundulaki, V. Christophides, and P. Boncz. Heuristics-based query optimisation for SPARQL. In *EDBT*, 2012.

[15] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *SIGKDD*, 2003.