Extending LargeRDFBench for Multi-Source Data at Scale for SPARQL Endpoint Federation

Ali HASNAIN\textsuperscript{a,1}, Muhammad SALEEM\textsuperscript{b} and Axel-Cyrille Ngonga NGOMO\textsuperscript{b,c} and Dietrich REBHOLZ-SCHUHMANN\textsuperscript{a}

\textsuperscript{a}Insight Centre for Data Analytics, National University of Ireland, Galway
\textsuperscript{b}Universit"at Leipzig, IFI/AKSW, PO 100920, D-04009 Leipzig
\textsuperscript{c}DICE, University of Paderborn, Germany

\textbf{Abstract.} Querying the Web of Data is highly motivated by the use of federation approaches mainly SPARQL query federation when the data is available through endpoints. Different benchmarks have been proposed to exploit the full potential of SPARQL query federation approaches in real world scenarios with their limitations in size and complexity. Previously, we introduced LargeRDFBench - a billion-triple benchmark for SPARQL query federation. In this work, we pinpoint some of the limitations of LargeRDFBench and propose an extension with 8 additional queries. Our evaluation results of the state-of-the-art federation engines revealed interesting insights, when tested on these additional queries.

\textbf{Keywords.} LargeRDFBench, Benchmarking, Federated SPARQL queries, RDF

1. Introduction

Due to linked, autonomous, and decentralised architecture of Linked Open Data (LOD), several queries require collecting information from more than one dataset also called data sources [1,2]. Processing such queries called federated queries are of central importance for the scale-able deployment of Semantic Web technologies. The importance of federated SPARQL queries for Linked Data management has led to the development of several federated SPARQL querying federation engines [3,6,7,5,13,12] etc. Consequently, this has motivated the design of several federated SPARQL querying benchmarks [9,10,11]. LargeRDFBench [9] addressed several limitation of FedBench [10] and Splodge [11]. In this work, we highlight some of the limitations of LargeRDFBench. In particular, the number of distinct datasets (sources for short) required to get the complete result set of the query is smaller in number (range between 1-4). As such, federation engines (e.g.,

\textsuperscript{1}Corresponding Author: Ali Hasnain, Insight Centre for Data Analytics, National University of Ireland, Galway; E-mail: ali.hasnain@insight-centre.org.
which optimise the ordering of the required distinct sources explicitly mentioned as SPARQL SERVICES cannot be fully tested with existing LargeRDFBench queries. To fill this gap, we extended the LargeRDFBench with 8 additional queries of varying complexities and number of distinct sources required. We discussed the key characteristics of each of these additional queries and evaluated state-of-the-art engines on these queries. The evaluation results revealed interesting insights about the performance and stability of these engines. The LargeRDFBench along with the proposed extension is available at: https://github.com/AKSW/largerdfbench.

2. Related Work

Over the last decade, various benchmarks have been proposed for comparing triple stores and SPARQL query processing systems. In this work, we only focus on federated SPARQL queries benchmarks. SPLODGE [11] benchmark uses heuristic for automatic generation of federated queries with conjunctive BGP. Non-conjunctive queries that make use of the SPARQL UNION, OPTIONAL clauses are not considered. FedBench [10] comprise of 9 real-world datasets and a total of 25 queries from different domains. Some of the limitations of FedBench was addressed in LargeRDFBench [9] with more real-world datasets and more complex and large data queries. In this work, we addressed some of the key limitations of LargeRDFBench and proposed an extension to this benchmark.

3. Design Features

In this section, we present the key SPARQL query features that should be considered while designing a federated SPARQL benchmark. Note that all of these key SPARQL features are formally presented in LargeRDFBench [9]. Here, we are re-introducing all of them for the sake of self containment of this paper and understanding the subsequent analysis.

The previous research contributions [11,9] on SPARQL querying benchmarking pointed out that SPARQL queries used in the benchmark should vary with respect to the the following key query characteristics: total number of triple patterns, number of join vertices, mean join vertex degree, number of sources span, query result set sizes, mean triple pattern selectivities, BGP-restricted triple pattern selectivity, join-restricted triple pattern selectivity, join vertex types (‘star’, ‘path’, ‘hybrid’, ‘sink’), and important SPARQL clauses used (e.g., LIMIT, OPTIONAL, UNION, FILTER etc.).

We represent any basic graph pattern (BGP) of a given SPARQL query as a directed hypergraph (DH) [14], a generalisation of a directed graph in which a hyperedge can join any number of vertices. In our specific case, every hyperedge captures a triple pattern. The subject of the triple becomes the source vertex of a hyperedge and the predicate and object of the triple pattern become the target vertices. For instance, the query (Figure 1) shows the hypergraph representation of a SPARQL query. Unlike a common SPARQL representation where the subject and object of the triple pattern are connected by an edge, our hypergraph-based representation contains nodes for all three components of the triple patterns. As a result, we can capture joins that involve predicates of triple patterns. Formally, our hypergraph representation is defined as follows:
A. Hasnain et al.

```sparql
SELECT DISTINCT *
WHERE {
    ?drug :drugType :smallMolecule.
    ?chemReaction :xEnzyme ?enzyme.
    ?chemReaction :equation ?chemEquation
    .
    ?chemReaction :title ?reactionTitle
}
```

**Figure 1.** DH representation of the SPARQL query

**Definition 3.1 (Directed hypergraph of a BGP)** The hypergraph representation of a BGP $B$ is a directed hypergraph $HG = (V, E)$ whose vertices are all the components of all triple patterns in $B$, i.e., $V = \bigcup_{(s,p,o) \in B} \{s, p, o\}$, and that contains a hyperedge $(S, T) \in E$ for every triple pattern $(s,p,o) \in B$ such that $S = \{s\}$ and $T = \{p, o\}$.

The representation of a complete SPARQL query as a DH is the union of the representations of the query’s BGPs. Based on the DH representation of SPARQL queries, we can define the following features of SPARQL queries:

**Definition 3.2 (Join Vertex)** For every vertex $v \in V$ in such a hypergraph we write $E_{in}(v)$ and $E_{out}(v)$ to denote the set of incoming and outgoing edges, respectively; i.e., $E_{in}(v) = \{(S, T) \in E \mid v \in T\}$ and $E_{out}(v) = \{(S, T) \in E \mid v \in S\}$. If $|E_{in}(v)| + |E_{out}(v)| > 1$, we call $v$ a join vertex.

**Definition 3.3 (Join Vertex Types)** A vertex $v \in V$ can be of type “star”, “path”, “hybrid”, or “sink” if this vertex participates in at least one join. A “star” vertex has more than one outgoing edge and no incoming edges. A “path” vertex has exactly one incoming and outgoing edge. A “hybrid” vertex has either more than one incoming and at least one outgoing edge or more than one outgoing and at least one incoming edge. A “sink” vertex has more than one incoming edge and no outgoing edge. A vertex that does not participate in joins is “simple”.

**Definition 3.4 (Number of Join Vertices)** Let $ST = \{st_1, \ldots, st_j\}$ be the set of vertices of type ‘star’, $PT = \{pt_1, \ldots, pt_k\}$ be the set of vertices of type ‘path’, $HB = \{hb_1, \ldots, hb_l\}$ be the set of vertices of type ‘hybrid’, and $SN = \{sn_1, \ldots, sn_m\}$ be the set of vertices of type ‘sink’ in a DH representation of a query, then the number of join vertices in the query $#JV = |ST| + |PT| + |HB| + |SN|$.

The total number of join vertices in a query is the sum of the total number of join vertices across all of the BGPs contained in this query.

**Definition 3.5 (Join Vertex Degree)** The DH representation of SPARQL queries makes use of the notion of $E_{in}(v) \subseteq E$ and $E_{out}(v) \subseteq E$ to denote the set of incoming and outgoing hyperedges of a vertex $v$. The joint vertex degree of a vertex $v$ is denoted $JVD_v = |E_{in}(v)| + |E_{out}(v)|$. 
The join vertex degree of the complete query is the average of all join vertex degrees of all the joins contained in this query. In our example (see Figure 1), the number of triple patterns is seven and the number of join vertices is four (two star, one sink and path each). The join vertex degree of each of the 'star' join vertex (shown in green colour) given in Figure 1 is three (i.e., three outgoing hyperedges from both vertices).

**Definition 3.6 (Relevant Source Set)** Let $D$ be the set of all data sources (e.g., SPARQL endpoints), $TP$ be the set of all triple patterns in query $Q$. Then, a source $d \in D$, is relevant (also called capable) for a triple pattern $tp_i \in TP$ if at least one triple contained in $d$ matches $tp_i$. The relevant source set $R_i \subseteq D$ for $tp_i$ is the set that contains all sources that are relevant for that particular triple pattern.

**Definition 3.7 (Total Triple Pattern-wise Sources)** By using Definition 3.6, we can define the total number of triple pattern-wise sources selected for query $Q$ as the sum of the magnitudes of relevant source sets $R_i$ over all individual triple patterns $tp_i \in Q$.

**Definition 3.8 (BGP-Restricted Triple Pattern Selectivity)** Consider a Basic Graph Pattern $BGP$ and a triple pattern $tp_i$ belonging to $BGP$, let $R(tp_i,D)$ be the set of distinct solution mappings (i.e., resultset) of executing $tp_i$ over dataset $D$ and $R(BGP,D)$ be the set of distinct solution mappings of executing $BGP$ over dataset $D$. Then the $BGP$-restricted triple pattern selectivity denoted by $Sel_{BGP-Restricted}(tp_i,D)$ is the fraction of distinct solution mappings in $R(tp_i,D)$ that are compatible (as per standard SPARQL semantics [15]) with a solution mapping in $R(BGP,D)$ [16]. Formally, if $\Omega$ and $\Omega'$ denote the sets underlying the (bag) query results $R(tp_i,D)$ and $R(BGP,D)$, respectively, then

$$Sel_{BGP-Restricted}(tp_i,D) = \left| \left\{ \mu \in \Omega \mid \exists \mu' \in \Omega' : \mu \text{ and } \mu' \text{ are compatible} \right\} \right| / |\Omega|$$

**Definition 3.9 (Join-Restricted Triple Pattern Selectivity)** Consider a join vertex $x$ in the DH representation of a $BGP$. Let $BGP'$ belonging to $BGP$ be the set of triple patterns that are incidents to $x$. Furthermore, let $tp_i$ belonging to $BGP'$ be a triple pattern and $R(tp_i,D)$ be the set of distinct solution mappings of executing $tp_i$ over dataset $D$ and $R(BGP',D)$ be the set of distinct solution mappings of executing $BGP'$ over dataset $D$. Then the $x$ - restricted triple pattern selectivity denoted by $Sel_{JVx-Restricted}(tp_i,D)$, is the fraction of distinct solution mappings in $R(tp_i,D)$ that are compatible with a solution mapping in $R(BGP',D)$ [16]. Formally, if $\Omega$ and $\Omega'$ denote the sets underlying the (bag) query results $R(tp_i,D)$ and $R(BGP',D)$, respectively, then

$$Sel_{JVx-Restricted}(tp_i,D) = \left| \left\{ \mu \in \Omega \mid \exists \mu' \in \Omega' : \mu \text{ and } \mu' \text{ are compatible} \right\} \right| / |\Omega|$$

### 4. Analysis

We now present analysis of the queries included in the original and the extended LargeRDFBench, based on the important query features introduced in the previous section.

\footnote{The concept of matching a triple pattern is defined formally in the SPARQL specification found at http://www.w3.org/TR/rdf-sparql-query/}
4.1. LargeRDBench

The original LargeRDBench comprises a total of 32 queries which are divided into three different types namely Simple, Complex, and Large Data queries. The Simple queries category includes a total of 14 queries (namely S1-S14). The Complex queries category includes a total of 10 queries (namely C1-C10). The Large Data queries category includes a total of 8 queries (namely L1-L8). Table 1 shows the characteristics of each query across the important query features discussed in the previous section. These queries were defined by considering increasing number of source selected per query. A brief summary of each queries category is given below.

4.1.1. Simple Queries

Simple queries were taken directly from the FedBench queries [10]. These queries are relatively fast to execute (around 2 seconds [9]) and include the smallest (in comparison to other categories) number of triple patterns. The number of triple patterns in this category range from 2 to 7. The number of join vertices and the mean join vertex degree for these queries are lower (average #JV = 2.6, MJVD = 2.1, ref. Table 1). Moreover, they only use a subset of the SPARQL clauses as shown in Table 1. Amongst others, they do not use LIMIT, REGEX, DISTINCT and ORDER BY clauses.

4.1.2. Complex Queries

The complex queries were particularly designed to address the aforementioned limitations of the simple queries. In particular, this queries category tackles the limitations with respect to the number of triple patterns, the number of join vertices, the mean join vertices degree, the SPARQL clauses, and the small query execution times of simple queries. Consequently, queries in this category rely on at least 8 triple patterns, i.e., one more than the maximum number (i.e. 7) of triple patterns in a simple query. The number of join vertices ranges from 3 to 6 (average #JV = 4.3, ref. Table 1). The mean join vertices degree ranges from 2 to 6 (average MJVD = 2.93, ref. Table 1). In addition, they were designed to use more SPARQL clauses, especially, DISTINCT, LIMIT, FILTER and ORDER BY. The evaluation results presented in LargeRDFBench show that the query execution time for complex queries exceeds 10 minutes.

4.1.3. Large Data Queries

The goal of the queries included in this category was to test the federation engines for real large data use cases. These queries span over large datasets and involve processing large intermediate result sets (usually in hundreds of thousands, see mean triple pattern selectivities in Table 1) or lead to large result sets (minimum 80459, see Table 1). The evaluation results presented in LargeRDFBench show that the query processing time for large data queries exceeds one hour.

4.2. Extended LargeRDFBench

One of the important question is to know the main motivation behind the need of an extension of LargeRDFBench with additional queries. The values inside the brackets of #RS column of Table 1 reveal one of the key problems with original LargeRDFBench
A. Hasnain et al.

Table 1. LargeRDFBench query characteristics. \(\#TP\) = total number of triple patterns, \(\#RS\) = distinct number of relevant source. The values inside brackets show the minimum number of distinct sources required to get the complete result set, \(\#JV\) = total number join vertices, MJVD = mean join vertices degree, MTPS = mean triple pattern selectivity, MBRTPS = Mean BGP-restricted triple pattern selectivity, MJRTPS = mean join-restricted triple pattern selectivity. UN = UNION, OP = OPTIONAL, DI = DISTINCT, FILTER, LIMIT, OB = ORDER BY, REGEX, NA = not applicable since there is no join node in the query, = no SPARQL clause used. STar, Sk, Path, HYbrid

<table>
<thead>
<tr>
<th>Query</th>
<th>Join Vertices</th>
<th>#TP</th>
<th>#RS</th>
<th>#Results</th>
<th>#JV</th>
<th>MJVD</th>
<th>MTPS</th>
<th>MBRTPS</th>
<th>MJRTPS</th>
<th>Clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1 ST</td>
<td>3</td>
<td>2</td>
<td>90</td>
<td>1</td>
<td>2</td>
<td>0.333</td>
<td>0.66667</td>
<td>0.33333</td>
<td>UN</td>
</tr>
<tr>
<td>S2</td>
<td>1 ST,1 P</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0.007</td>
<td>0.66671</td>
<td>0.33333</td>
<td>-</td>
</tr>
<tr>
<td>S3</td>
<td>1 ST,1 HY</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0.008</td>
<td>0.00031</td>
<td>0.00015</td>
<td>-</td>
</tr>
<tr>
<td>S4</td>
<td>2 ST,2 SI,1 P</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>0.019</td>
<td>0.20030</td>
<td>0.20003</td>
<td>-</td>
</tr>
<tr>
<td>S5</td>
<td>1 ST,2 P</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0.006</td>
<td>0.00067</td>
<td>0.00064</td>
<td>-</td>
</tr>
<tr>
<td>S6</td>
<td>1 ST,2 P</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>11</td>
<td>3</td>
<td>0.019</td>
<td>0.25013</td>
<td>0.25001</td>
<td>-</td>
</tr>
<tr>
<td>S7</td>
<td>1 ST,2 P</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0.020</td>
<td>0.00227</td>
<td>0.00227</td>
<td>-</td>
</tr>
<tr>
<td>S8</td>
<td>No Join</td>
<td>2</td>
<td>1</td>
<td>1159</td>
<td>0</td>
<td>NA</td>
<td>0.001</td>
<td>1.00000</td>
<td>0.00000</td>
<td>UN</td>
</tr>
<tr>
<td>S9</td>
<td>1 P</td>
<td>3</td>
<td>13</td>
<td>333</td>
<td>1</td>
<td>2</td>
<td>0.333</td>
<td>0.37037</td>
<td>0.03704</td>
<td>UN</td>
</tr>
<tr>
<td>S10</td>
<td>1 ST,2 P</td>
<td>5</td>
<td>2</td>
<td>9054</td>
<td>3</td>
<td>2.33</td>
<td>0.016</td>
<td>0.55678</td>
<td>0.00011</td>
<td>-</td>
</tr>
<tr>
<td>S11</td>
<td>2 ST,1 SI,1 HY</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2.5</td>
<td>0.006</td>
<td>0.04800</td>
<td>0.04788</td>
<td>-</td>
</tr>
<tr>
<td>S12</td>
<td>2 ST,1 P,1 SI</td>
<td>6</td>
<td>5</td>
<td>393</td>
<td>4</td>
<td>2.25</td>
<td>0.012</td>
<td>0.07563</td>
<td>0.00020</td>
<td>-</td>
</tr>
<tr>
<td>S13</td>
<td>3 ST</td>
<td>5</td>
<td>5</td>
<td>28</td>
<td>3</td>
<td>2.33</td>
<td>0.014</td>
<td>0.11996</td>
<td>0.00428</td>
<td>-</td>
</tr>
<tr>
<td>S14</td>
<td>2 ST,1 SI</td>
<td>5</td>
<td>3</td>
<td>1620</td>
<td>3</td>
<td>2</td>
<td>0.012</td>
<td>0.48156</td>
<td>0.00026</td>
<td>OP</td>
</tr>
</tbody>
</table>

Avg. 4.3 5.7(2.1) 907 2.6 2.1 0.057 0.31713 0.08640

| Avg. | 9.2 7.8(2.4) 509 2.4 2.3 0.013 0.22127 0.01195
| Avg. | 7.62 4.6(2.1) 236793 3.75 2.70 0.293 0.29092 0.00998
| CH1  | 2 ST, 1 P, 1 SI, 1 HY | 16  | 8     | 384      | 5    | 4.2  | 0.00200| 0.00099| 0.1256 | DI, OB |
| CH2  | 4 ST, 1 P, 1 SI, 0 HY | 10  | 9     | 840      | 6    | 2.17 | 0.00005| 0.2595 | 0.58115| DI, FI |
| CH3  | 2 ST, 1 P, 3 SI, 1 HY | 11  | 13    | 185757   | 5    | 3   | 0.485| 0.39644| 0.00367| FI      |
| CH4  | 4 ST, 3 P, 2 SI, 1 HY | 12  | 8     | 1248     | 10   | 2.3 | 0.00074| 0.1184 | 0.2586 | DI, FI |
| CH5  | 4 ST, 2 P, 2 SI, 2 HY | 18  | 13    | 755      | 10   | 2.6 | 0.05561| 0.2946 | 0.3945 | DI, LI |
| CH6  | 5 ST, 3 P, 2 SI, 2 HY | 24  | 10    | 16       | 12   | 2.83| 0.00004| 0.2522 | 0.3186 | -       |
| CH7  | 5 ST, 3 P, 4 SI, 2 HY | 21  | 12    | 775      | 14   | 2.42| 0.00007| 0.2135 | 0.2754 | DI, LI |
| CH8  | 7 ST, 4 P, 6 SI, 2 HY | 33  | 13    | 19       | 2.63 | 0.00073| 0.2354 | 0.2546 | OP, FI, LI |

queries. Note that these values show the minimum number of distinct sources required to get the complete result set. Furthermore, these values also show the total number of distinct SERVICES used in the SPARQL 1.1 version of each of the benchmark queries. By looking at Table 1, majority of the simple queries (except S6, S12) require only 2 distinct sources to get the complete result set of the queries. Even in the complex queries
category (i.e., C1-C10), the maximum number of distinct sources to get the complete result of the query is only 3. Finally, in the large data queries category (i.e., L1-L8), there is only one query which requires 3 distinct sources. All others in this category require only two data sources.

As an overall result, there are 24 out of total 32 queries which only require 2 data sources to get the complete result set of the queries. This clearly shows that the federation engines (e.g., [8]) which optimise the ordering of the execution of SPARQL services in federated SPARQL 1.1 queries cannot be fully tested with existing LargeRDFBench queries. This is because if there are only two services used in the query, there are only two possible orderings of the execution of these services. As such, even the probability of random service ordering is 0.5 without the need of any heuristics or cost model.

The goal of this extension was to fill this gap by adding more federated queries which require more data sources to get the complete result set of the query. We now describe the queries which are added into the LargeRDFBench.

4.2.1. Complex and High Data Sources Queries

In this work, we added 8 additional Complex and High data sources (namely CH1-CH8) queries into LargeRDFBench, making the total number of queries in the benchmark equal to 40. These queries have increasing numbers (from 4-10) of the distinct data sources required to get the complete result set. In addition, the number of join vertices and the number triple patterns in these queries are much higher than existing LargeRDFBench queries (see Table 1). Consequently, we will see in our evaluation that the query runtimes of these queries ranges from less than one second to more than 1 hour. All the extended queries are given at the end of the paper as in Appendix and their key characteristics are discussed below.

**CH1:** This query requires 4 LargeRDFBench data sources, i.e., DBpedia, New York Times, Geonames, and Semantic Web Dog Food to get the complete result set. The total triple patterns in this query is 16 with result size of 384. The key characteristic of query is the high mean join vertex degree (i.e., 4.2, ref. Table 1). This means that the number of incoming and outgoing edges of a join node is high compared to other queries in the benchmark. In other words, the number of triple patterns in a single service would be relatively high. A smart query planner can combine a set of triple patterns and send them all in to the relevant source as single group. Federation engine needs not to perform the join between triple patterns. Rather, the join can be migrated to the relevant source (i.e., the SPARQL endpoint) and hence can greatly improve the runtime performance by dividing the load between endpoints and federation engine. On the other hand, a good estimation of the cardinality of multi-triple patterns join can be particularly challenging. A wrong estimation, can leads to a wrong query execution plan. We will see in our evaluation results, the runtime for this query ranges from 1 second to over 1 hour for the different federation engines.

**CH2:** This query requires 4 LargeRDFBench data sources, i.e., DBpedia, DrugBank, KEGG, and ChEBI to get the complete result set. The total triple patterns in this query is 10 with result size of 840. The key characteristic of this query is the low mean triple pat-

---

3 Please look at LargeRDFBench home page for the datasets included in the benchmark
terns selectivity (i.e., 0.00005) and high BGP-restricted (i.e., 0.2595) and Join-restricted (i.e., 0.58115) triple pattern selectivities. This means that the triple patterns of this query are selective, i.e., they can have smaller result sizes. However, they are less selective when involved in joins with other triple patterns. Since the triple patterns are selective, choosing the right join order which quickly converges to smaller result size is particularly crucial. The FILTER combined with REGEX made this query particularly very selective.

**CH3:** This query involves 5 data sources, i.e., DBpedia, DrugBank, KEGG, ChEBI, Linked TCGA-A to get the complete result set. The triple patterns involved in this query is 11 with result size of 48. Query has relatively high number of join vertices (i.e., 7 from 11 triple patterns), moderate join vertex degree (i.e., 2.71), and low mean BGP-restricted triple pattern selectivity (i.e., 0.0196). This query is a candidate example of using a mix values for the important query features and can challenge the federation engines for a mix of these values.

**CH4:** This query requires 6 LargeRDFBench data sources, i.e., DBpedia, Semantic Web Dog Food, GeoNames, New York Times, Jamendo, and Linked MDB to get the complete result set. The total number of triple patterns in this query is 12 with result size of 1248. This query has relatively very high number of join vertices (i.e., 10 join vertices from 12 triple patterns). This means that the join order optimisation of this query can be particularly challenging due to more joins with less number of triple patterns involved in the joins.

**CH5:** This query requires 7 data sources, i.e., DBpedia, Linked TCGA-A, DrugBank, GeoNames, New York Times, Jamendo, and Linked MDB to receive the complete result set. The triple patterns involved in this query is 18 with result size of 5 using LIMIT clause. The triple patterns in aforementioned queries where mostly unbound subjects and objects with bound predicate. Unlike the previous queries, the key characteristic of this query is number of bounds subjects and objects in the triple patterns. There are 4 triple patterns for which the subject is bound and 2 triple patterns for which object is bound. Also there is one triple pattern that contains unbound predicate. Thus, this query can be challenging to accurately estimate the triple patterns as well as the joins cardinalities due to bound subjects and objects as well as unbound predicate in triple patterns.

**CH6:** This query requires 8 LargeRDFBench data sources, i.e., Linked TCGA-A, DBpedia, DrugBank, KEGG, GeoNames, New York Times, Jamendo, and Linked MDB to get the complete result set. The triple patterns of this query is 24 with result size of 16. Similar to CH2, The key characteristic of this query is the low mean triple patterns selectivity (i.e., 0.00002) and high BGP-restricted (i.e., 0.2522) and Join-restricted (i.e., 0.3186) triple pattern selectivities. This query also contains triple patterns with bound subjects and objects.

**CH7:** This query requires 9 LargeRDFBench data sources, i.e., Linked TCGA-A, DBpedia, DrugBank, KEGG, GeoNames, Semantic Web Dog Food, New York Times, Jamendo, and Linked MDB to get the complete result set. The total number of triple patterns in this query is 21 with result size of 775 using LIMIT clause. There are a total of 14 join nodes in this query with 5 Star, 3 Path, 4 Sink, and 2 Hybrid join nodes.
query be particularly challenging due to more join nodes and hence join ordering could not be a trivial task.

**CH8:** This query requires 10 data sources, i.e., Linked TCGA-A, DBpedia, DrugBank, KEGG, GeoNames, Semantic Web Dog Food, New York Times, Jamendo, ChEBI, and Linked MDB to get the complete result set. This query contains the highest number of triple patterns among the benchmark queries (i.e., 33). The result size of this query is only 1. There are a total of 19 join nodes in this query with 7 Star, 4 Path, 6 Sink, and 2 Hybrid join nodes. This query also contains OPTIONAL, FILTER, and LIMIT. As one of the most complex query of the benchmark our evaluation (section 5) shows that non of the federation engines is able to execute this query within the timeout limit of 1 hour.

5. Evaluation

In this section, we evaluate state-of-the-art SPARQL query federation systems by using the extended queries added into the LargeRDFBench. We first describe our experimental setup in detail. Then, we present our evaluation results. All data used in this evaluation can be found on the benchmark homepage.

5.1. Experimental Setup

The experimental setup was used as of the original LargeRDFBench evaluation. In summary, LargeRDFBench contains a total of 13 real-world datasets. Each of the datasets were loaded into a Virtuoso 7.1. Each of the 13 Virtuoso SPARQL endpoints used in our experiments was installed on a separate machine. The specification of each of the machines are exactly the same used in the original LargeRDFBench. We ran the extended queries experiments on a clustered server with 32 physical CPU cores of 2.10GHz each and a total RAM of 512GB. Each of the 13 Virtuoso SPARQL endpoints used in our experiments was started as a separate instance on the clustered server. The federation engines were also run on the same machine. We set the maximal amount of memory for each of the federation engines to 128GB. Experiments were conducted on local copies of Virtuoso SPARQL endpoints with number of buffers 1360000, maximum dirty buffers 1000000, number of server threads 20, result set maximum rows 100,000,000,000 and maximum SPARQL endpoint query execution time of 6000,000,000 seconds. The query timeout was set 1 hour. Seven SPARQL endpoint federation engines (versions available as of May 2018) were compared – FedX [7], SPLENDID [5], ANAPSID [5], FedX+HiBISCuS [14], SPLENDID+HiBISCuS [14], SemaGrow [4], CostFed [3] – on all of the 8 extended benchmark queries. We used all of the performance metrics used in the original LargeRDFBench except for the number of endpoints requests which is proven to not necessarily correlate with the overall query runtimes [9]. We used: (1) the total number of triple pattern-wise (TPW) sources selected during the source selection, (2) the total number of SPARQL ASK requests submitted to perform (1), (3) the completeness (recall) and correctness (precision) of the query result set retrieved, (4) the average source selection time, (5) the average query execution time.
Table 2. Comparison of the different SPARQL federation engines in terms of: (1) Total number of triple pattern-wise sources selected #TP and (2) Total SPARQL ASK request #AR, where H-FedX = HiBISCuS+FedX, H-SPLENDID = HiBISCuS+SPLENDID, ZR = Zero Results, RE = Runtime Error, PE = Parse Error.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CH1</td>
<td>41</td>
<td>208</td>
<td>22</td>
<td>0</td>
<td>41</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>CH2</td>
<td>20</td>
<td>130</td>
<td>11</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>CH3</td>
<td>37</td>
<td>143</td>
<td>19</td>
<td>0</td>
<td>37</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>CH4</td>
<td>41</td>
<td>169</td>
<td>13</td>
<td>0</td>
<td>41</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>CH5</td>
<td>57</td>
<td>234</td>
<td>22</td>
<td>0</td>
<td>57</td>
<td>78</td>
<td>22</td>
</tr>
<tr>
<td>CH6</td>
<td>40</td>
<td>312</td>
<td>RE</td>
<td>RE</td>
<td>40</td>
<td>104</td>
<td>RE</td>
</tr>
<tr>
<td>CH7</td>
<td>52</td>
<td>273</td>
<td>28</td>
<td>0</td>
<td>52</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>CH8</td>
<td>68</td>
<td>403</td>
<td>38</td>
<td>0</td>
<td>68</td>
<td>13</td>
<td>38</td>
</tr>
</tbody>
</table>

5.2. Experimental Results

5.2.1. Efficiency of Source Selection

Similar to LargeRDFBench, we define efficient source selection in terms of: (1) the total number of triple pattern-wise sources selected (#T), (2) the total number of SPARQL ASK requests (#AR) used to obtain (1), and (3) the source selection time (SST). Table 2 shows the results of these three metrics for the selected approaches.

Overall, CostFed (total 151 #TP, ref. Table 2) is the most efficient approach in terms of smaller number of total TPW sources selected, followed by HiBISCuS (total 153 #TP) and ANAPSID (total 161 #TP). Which is equally followed by FedX, SPLENDID, and SemaGrow with 356 #TP each. Interestingly, ANAPSID which was the best approach in terms of #TP for original LargeRDFBench ranked third in our proposed extension. The reason behind this is as the number of triple patterns increases in the queries, the efficient source selection becomes more difficult. The SSGM heuristics [6] used in the ANAPSID may not work that efficient with increasing number of triple patterns. CostFed, HiBISCuS, and ANAPSID are equally best approaches in terms of smaller number of ASK requests used (i.e., 0 for all queries). Noteworthy, FedX (cold) (without cache) used a total of 1872 ASK requests. This is because FedX(cold) needs to sent a request for each triple pattern to each of the 13 SPARQL endpoints (hosted on separate machines). In terms of the source selection time, FedX (warm) (with cache) is the fastest approach (avg. 6 msec, ref, Table 3) followed by CostFed (avg. 39 msec), SemaGrow (avg. 54 msec), HiBISCuS (avg. 195 msec), FedX (cold) (avg. 239 msec), and ANAPSID (avg. 322 msec). The results clearly suggest that the FedX source selection is greatly improved by using caching.

5.2.2. Query Runtime

Table 3 (column RT) shows the query runtime performances of the selected federation engines. As an overall performance evaluation, it is rather hard to rank the selected engines as there are many timeouts, runtime or parse errors, suggesting the selected federation engines are not that stable when tested with queries containing more triple patterns and require collecting results from more sources (i.e., greater than 3). CostFed has the smallest query runtime for CH1 (i.e., 800 msec) while the same query time out for ANAPSID. For CH2, ANAPSID has the smallest query runtime (i.e., 101 min) while the
Table 3. SPARQL federation engines comparison in terms of: (1) Total number of endpoints requests \#EPR, (2) Source selection time (cold) SSTC, (3) Source selection time (warm) SSTW, (4) RT = Runtime (msec) cold cache, (5) RTW =Runtime (msec) warm cache, (6) RT = Runtime (msec), where H-FedX = HiBiCuS+FedX, H-SPLENDID = HiBiCuS+SPLENDID, ZR = Zero Results, RE = Runtime Error, PE = Parse Error, TO = TimeOut 1 hour.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CH1</td>
<td>144</td>
<td>6</td>
<td>6348</td>
<td>6321</td>
<td>124</td>
<td>4109</td>
<td>124</td>
<td>4109</td>
</tr>
<tr>
<td>CH2</td>
<td>240</td>
<td>5</td>
<td>328265</td>
<td>3282233</td>
<td>135</td>
<td>778563</td>
<td>10</td>
<td>137</td>
</tr>
<tr>
<td>CH3</td>
<td>178</td>
<td>7</td>
<td>332573</td>
<td>275412</td>
<td>92</td>
<td>329223</td>
<td>8</td>
<td>88</td>
</tr>
<tr>
<td>CH4</td>
<td>107</td>
<td>4</td>
<td>ZR</td>
<td>ZR</td>
<td>247</td>
<td>ZR</td>
<td>66</td>
<td>240</td>
</tr>
<tr>
<td>CH5</td>
<td>295</td>
<td>5</td>
<td>TO</td>
<td>TO</td>
<td>474</td>
<td>RE</td>
<td>134</td>
<td>24719</td>
</tr>
<tr>
<td>CH6</td>
<td>445</td>
<td>7</td>
<td>715</td>
<td>274</td>
<td>RE</td>
<td>154</td>
<td>15922</td>
<td>RE</td>
</tr>
<tr>
<td>CH7</td>
<td>264</td>
<td>5</td>
<td>TO</td>
<td>TO</td>
<td>96</td>
<td>RE</td>
<td>8</td>
<td>96</td>
</tr>
<tr>
<td>CH8</td>
<td>236</td>
<td>6</td>
<td>TO</td>
<td>TO</td>
<td>247</td>
<td>RE</td>
<td>37</td>
<td>231</td>
</tr>
</tbody>
</table>

same query almost timeout for FedX (i.e., 55 min). For CH3, SemaGrow is the fastest while both CostFed and ANAPSID gives zero results. Only ANAPSID and SemaGrow is able to get complete results for CH4. For CH5, HiBiCuS+SPLENDID is the fastest (i.e., 15 sec) while SemaGrow and FedX timeout. For CH6, CostFed is the fastest (i.e., only 173 msec) while SemaGrow timeout. Note that for this query FedX gives 1 result and CostFed gives 4 results while actual result size is 16. For CH7, ANAPSID is the fastest (i.e., 2.4 sec) while FedX, SemaGrow time out. All of the selected engines timeout of 1 hour for CH8.

In conclusion, above results clearly suggest that the extended LargeRDFBench can be extremely costly or can be executed extremely fast when proper optimised query plan is selected. However, the number of timeout and runtime errors suggesting that choosing the optimised query plans for these queries is not a trivial task. The results revealed FedX, CostFed, and ANAPSID can result in incomplete or zero results.

Acknowledgement

This work was supported by the project HOBBIT, which has received funding from the European Union’s H2020 research and innovation action program (GA number 688227). Also, this publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund.

References


[8] Yannakis Thanos, Fafalios Pavlos and Tzitzikas Yannis Heuristics-based Query Reordering for Federated Queries in SPARQL 1.1 and SPARQL-LD, QuWeDa at ESWC 2018


Appendix A: Queries

```sparql
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX tcga: <http://tcga.deri.ie/schema/>
PREFIX kegg: <http://bio2rdf.org/ns/kegg#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX geonames: <http://www.geonames.org/ontology#>
PREFIX linkedmdb: <http://data.linkedmdb.org/resource/movie/>
PREFIX linkedmdbr: <http://data.linkedmdb.org/resource/>
PREFIX purl: <http://purl.org/dc/elements/1.1/>
PREFIX bio2rdf: <http://bio2rdf.org/ns/bio2rdf#>
PREFIX chebi: <http://bio2rdf.org/ns/chebi#>
PREFIX swc: <http://data.semanticweb.org/ns/swc/ontology#>
PREFIX eswc: <http://data.semanticweb.org/conference/eswc/>
PREFIX swcp: <http://data.semanticweb.org/person/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/>
PREFIX drug: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/>

################### CH1 ###################

SELECT DISTINCT *
WHERE {
  ?place geonames:name ?countryName;
  geonames:countryCode ?countryCode;
  geonames:population ?population ;
  geo:long ?longitude ;
  geo:lat ?latitude ;
  owl:sameAs ?geonameplace.
  ?geonameplace dbpedia:capital ?capital ;
  dbpedia:anthem ?nationalAnthem ;
  dbpedia:foundingDate ?foundingDate ;
  dbpedia:largestCity ?largestCity ;
  dbpedia:ethnicGroup ?ethnicGroup ;
  dbpedia:motto ?motto .
  ?role swc:heldBy ?writer .
  ?nytimes:latest_use ?dateused }
ORDER BY DESC (?population)

LargeRDFBench (new) queries.
```
A. Hasnain et al.

SELECT DISTINCT ?drug ?drugBankName ?keggmass ?chebiIupacName
WHERE
{
?dbPediaDrug dbpedia:casNumber ?casNumber .
?chebiDrug chebi:iupacName ?chebiIupacName .
FILTER REGEX (?chebiIupacName, "adenosine")
}

SELECT *
WHERE
{
?chebiDrug purl:title ?drug ;
bio2RDF:mass ?mass ;
bio2RDF:formula ?formula ;
bio2RDF:urlImage ?image
}
ORDER BY (?mass)

SELECT DISTINCT *
WHERE
{
?role swc:isRoleAt eswc:2010.
?role swc:heldBy ?author .
?dbpediaCountry owl:sameAs ?geoname ;
nytimes:latest_use ?dateused ;
owl:sameAs ?geoname .
?artist foaf:based_near ?geoname ;
?mdbFilm owl:sameAs ?film .
FILTER REGEX (?geoname, "United")
}

LargeRDFBench (new) queries.
SELECT DISTINCT *
WHERE {
  ?country dbpedia:populationDensity 32 .
  ?nytimesCountry owl:sameAs ?country ;
  nytimes: latest_use ?dateused ;
  owl:sameAs ?geonames .
  ?artist foaf:based_near ?geoname ;
  foaf:homepage ?homepage .
  ?x owl:sameAs ?film .
  drug:DB00441 drug:genericName ?drugName .
  drug:DB00441 drugbank:keggCompoundId ?compound .
}
LIMIT 5

SELECT ?patient ?country ?articleCount ?chemicalStructure ?id
WHERE {
  ?nytimesCountry owl:sameAs ?country ;
  nytimes: latest_use ?latestused ;
  nytimes: number_of_variants ?totalVariants ;
  nytimes: associated_article_count ?articleCount ;
  owl:sameAs ?geonames .
  swcp: christian-bizer foaf:based_near ?geoname ;
  foaf:homepage ?homepage .
  dbr:The_Last_Valley dbpedia:director ?director .
  ?x owl:sameAs dbr:The_Last_Valley .
  ?drugbcr tcga:drug_name "Cisplatin" .
  ?drugBnkDrg drugbank:inchiKey ?inchiKey .
  ?drugBnkDrg drugbank:casRegistryNumber ?id .
  ?keggDrug bio2rdf:xRef ?id .
}
SELECT DISTINCT * WHERE {
?nytimesCountry owl:sameAs ?country ;
nytimes:search_api_query ?apiQuery ; owl:sameAs ?location .
?artist foaf:based_near ?location ; foaf:firstName ?firstName .
?x owl:sameAs ?film .
?drgBnkDrg drugbank:casRegistryNumber ?id .
?drgBnkDrg drugbank:brandName ?brandName .
}
LIMIT 775

SELECT * WHERE {
?nytimesCountry owl:sameAs ?country ; nytimes:latestUse ?latestused ;
nytimes:number_of_variants ?totalVariants ;
nytimes:associatedArticleCount ?articleCount ;
owl:sameAs ?geonames .
?role swc:isRoleAt eswc:2010 .
?role swc:heldBy ?author .
?x owl:sameAs ?film .
?drgBnkDrg drugbank:inchiKey ?inchiKey .
?keggDrug rdf:type kegg:Drug ;
bio2rdf:xRef ?id .
?chebiDrug chebi:iupacName ?chebiIupacName .
OPTIONAL {
FILTER (?drugbankInchi = ?chebiInchi) } } LIMIT 1

LargeRDFBench (new) queries.