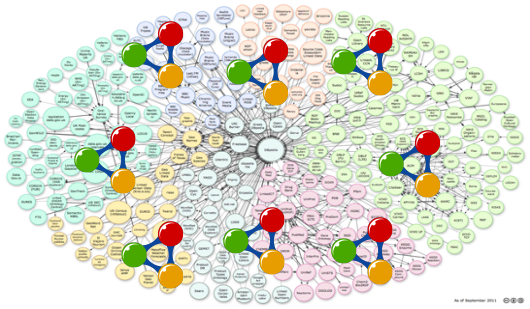
UNIVERSIDAD SIMÓN BOLÍVAR

SILURIAN: a Sparql vIsuaLizer for UndeRstanding querIes And federatioNs

Simón Castillo, Guillermo Palma and Maria-Esther Vidal
{scastillo,gpalma,mvidal}@ldc.usb.ve

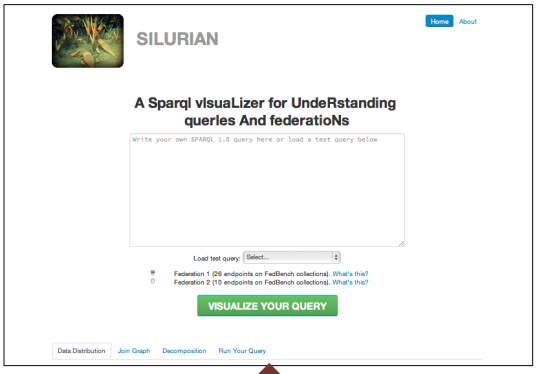
Motivation

Explosion on the number and size of Linked Datasets
- **SPARQL** Endpoints to access RDF data.
- **Applications** from different domains require to gather data from several endpoints.
- **Performance** of Federated SPARQL queries can be affected by diverse parameters, e.g., number of triple patterns, endpoints and shape of the query.



Goals:
- **Visualize** SPARQL **queries** and **federations** to understand the complexity of the different plans.

Approach



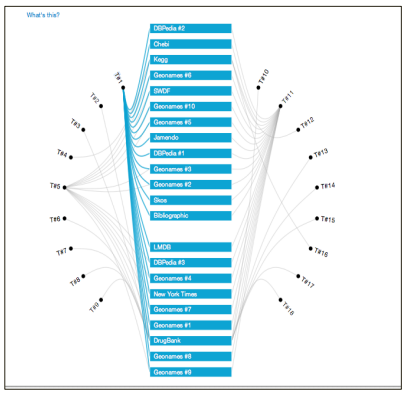
FedBench

Demonstration Use Cases

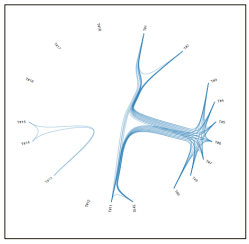
Query: Drugs that possibly target Leukemia

```
PREFIX: drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/>
PREFIX dbcategory:<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugcategory/>
PREFIX owl: <http://www.w3.org/2002/07/owl\#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns\#>
PREFIX dbowl:<http://dbpedia.org/ontology/>
PREFIX kegg: <http://bio2rdf.org/ns/kegg\#>
PREFIX diseasome: <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseases>
SELECT DISTINCT ?drug1
WHERE
{?drug1 drugbank:possibleDiseaseTarget diseasome:673 .
 ?drug1 drugbank:target ?o.
 ?o drugbank:genbankIdGene ?g.
 ?o drugbank:locus ?l.
 ?o drugbank:molecularWeight ?mw.
 ?o drugbank:hprdId ?hp.
 ?o drugbank:swissprotName ?sn.
 ?o drugbank:proteinSequence ?ps.
 ?o drugbank:generalReference ?gr.
 ?drug drugbank:target?o.
 ?drug drugbank:synonym?o1 .
 OPTIONAL { ?drug owl:sameAs ?drug5 .
            ?drug5 rdf:type dbcategory:Drug .
            ?drug drugbank:keggCompoundId ?cpd .
            ?enzyme kegg:xSubstrate ?cpd .
            ?enzyme rdf:type kegg:Enzyme .
           ?reaction kegg:xEnzyme ?enzyme .
           ?reaction kegg:equation ?equation . } }
```

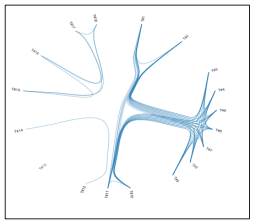
Endpoint Distribution Per Triple Patterns



Exclusive Groups Based Decomposition [1]



Star-Shaped Group Multiple Endpoint (SSGM) Decomposition[2]



Discussion

Triple patterns bound to general predicates, e.g., from RDFS or OWL,
- all **endpoints** may need to be contacted to produce complete answers.

Queries with **large number** of triple patterns may
- be decomposed in a large number of **sub-queries**, and
- require to execute **costly** sub-queries.

Data **partition** and **replication** may **negatively** impact on performance
- **Relevant** endpoints increase according to data fragments,
- **Execution** time may be affected by **vertical** fragmentation,
- **Completeness** may be impacted for **horizontal** fragmentation.

[1] Andreas Schwarte et al, "FedX: Optimization Techniques for Federated Query Processing on Linked Data". ISWC 2011.
[2] Gabriela Montoya et al, "A Heuristic-Based Approach for Planning Federated SPARQL Queries". COLD 2012.

Conclusions and Future Work

Summarizing,

- **SILURIAN** visualizes parameters that impact on the complexity of federated queries, e.g.,
  - ✓ data fragmentation and replication,
  - ✓ triple patterns bound to general predicates,
  - ✓ query shape, and
  - ✓ between answer completeness and execution time.

In the Future,
- **SILURIAN** will visualize plans produced by state-of-the-art federated engines,
- Extend **SILURIAN** to understand the impact of dynamicity and data updates.