HIBISCUS: Hypergraph-Based Source Selection for Federated SPARQL Queries

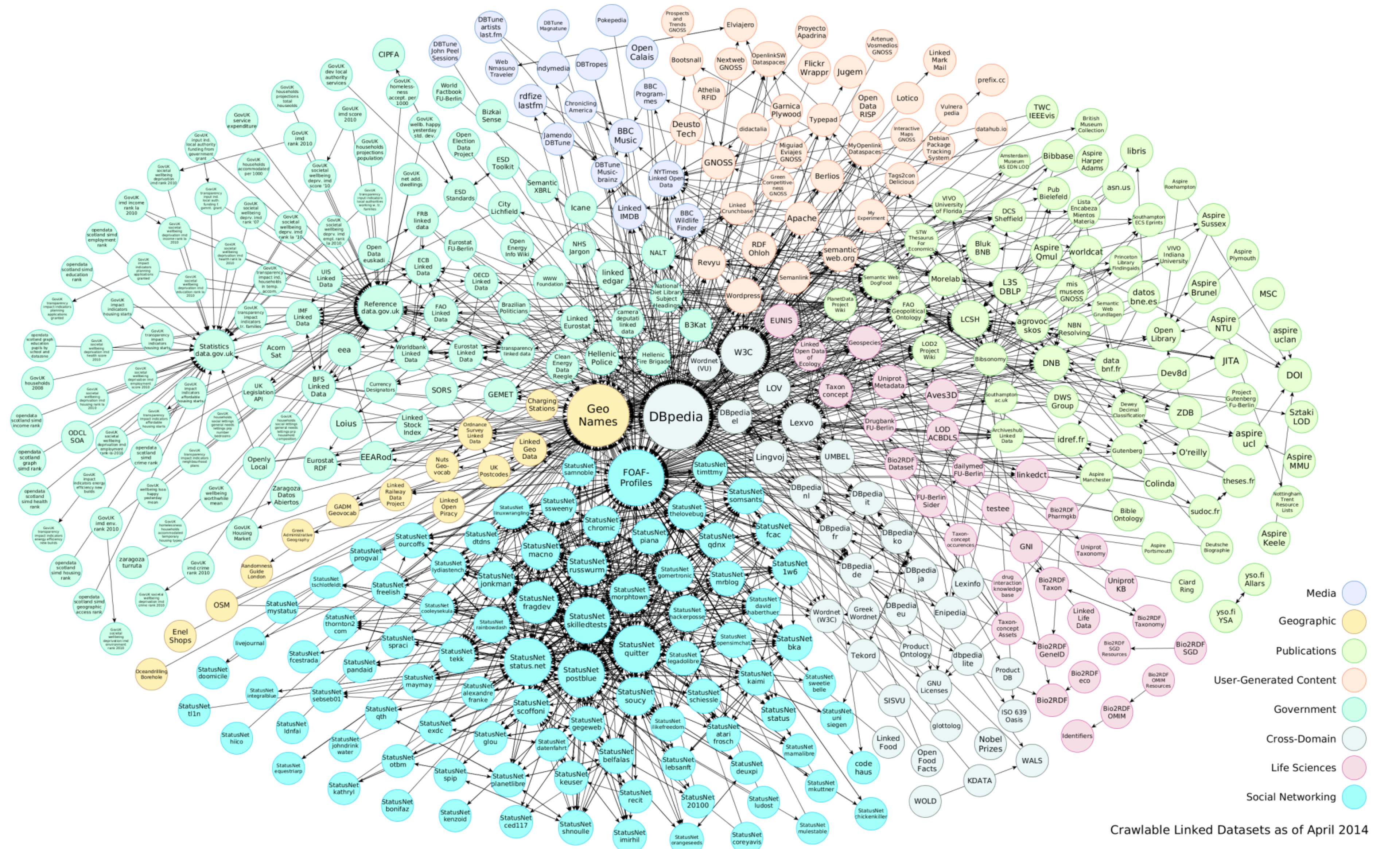
Axel-Cyrille Ngonga Ngomo
Joint work with Muhammad Saleem
AKSW Research Group

CREDIBLE Thematic Working Days

October 9th, 2014

Nice, France

Motivation



Media
Geographic
Publications
User-Generated Content
Government
Cross-Domain
Life Sciences
Social Networking
Crawlable Linked Datasets as of April 2014

Federation

- Linked Data Federation
- SPARQL Federation
- Hybrid Approaches

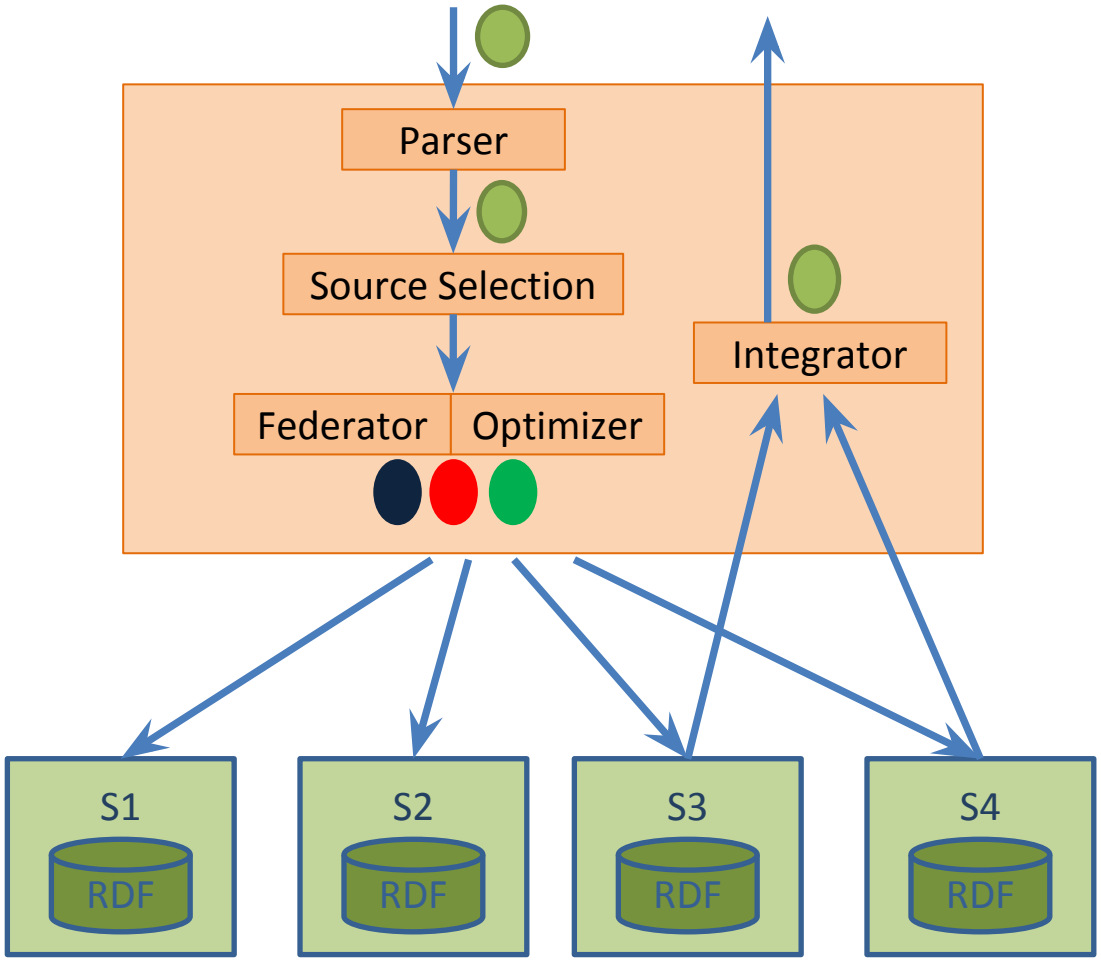
Federation

- Linked Data Federation
- **SPARQL Federation**
- Hybrid Approaches

FedBench (LD3): Return all US presidents, their party membership and news pages about them.

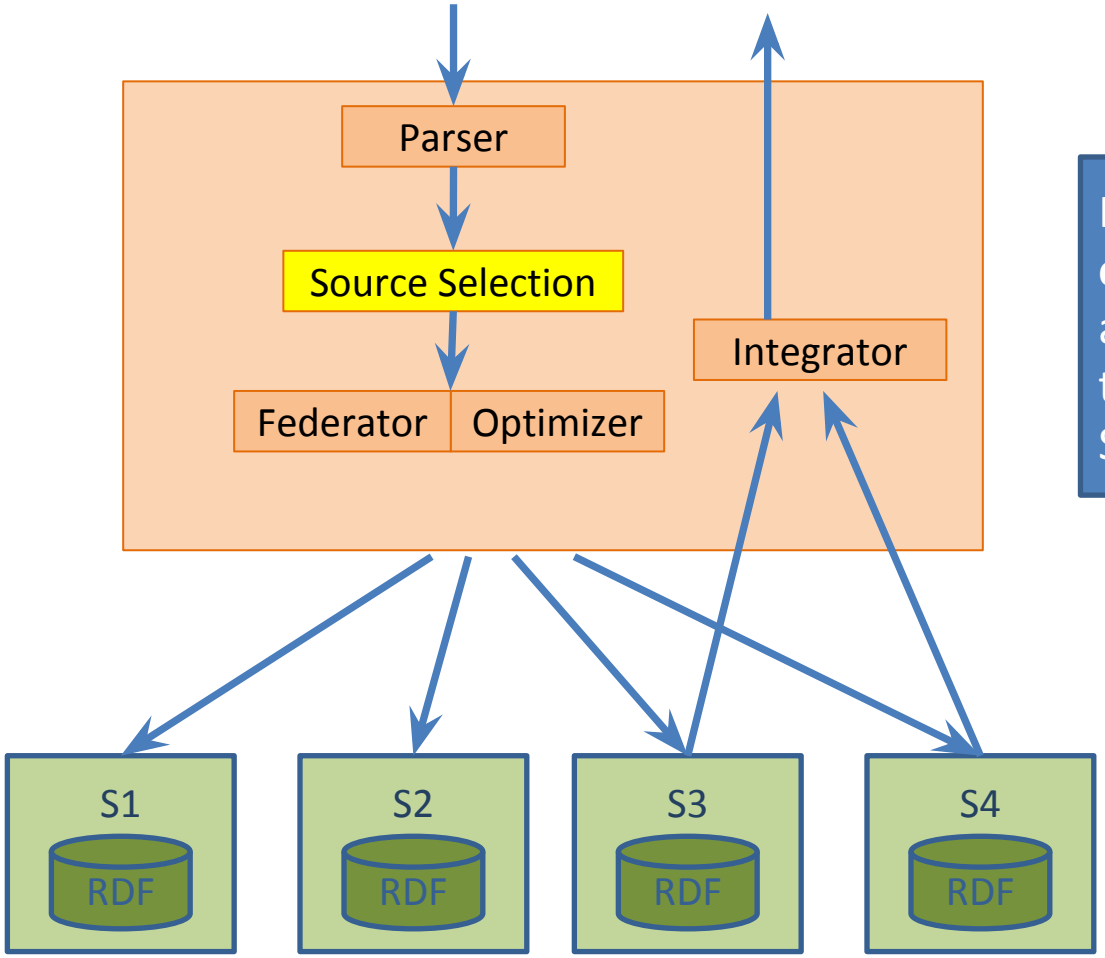
```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```

Federation



- Get individual triple patterns
- Identify capable sources against each triple pattern
- Generate query execution plan
- Execute sub-queries
- Integrate sub-queries results

HIBISCUS



Identify **capable and contributing** sources against individual triple patterns of a SPARQL query

Federation

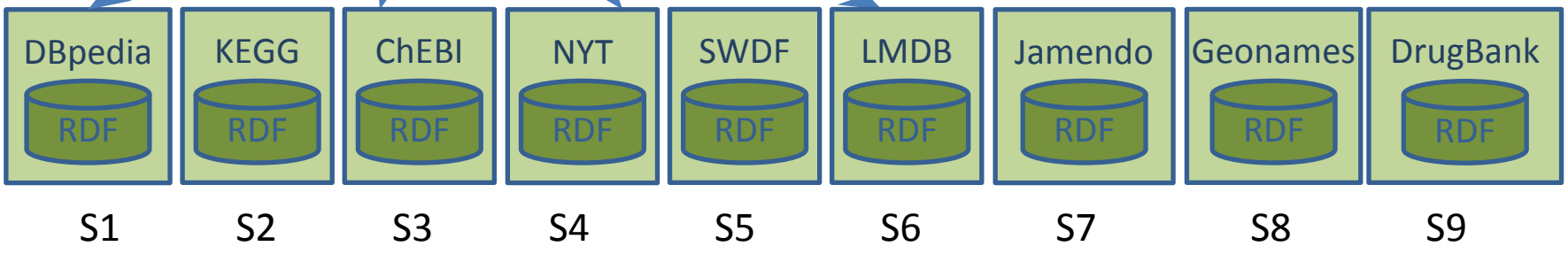
FedBench (LD3): Return all US presidents, their party membership and news pages about them.

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .      //TP1
?president dbpedia:nationality  dbpedia:United_States . //TP2
?president dbpedia:party ?party .     //TP3
?x  nyt:topicPage ?page .             //TP4
?x  owl:sameAs ?president .           //TP5
}
```

Source Selection Algorithm

Triple pattern-wise source selection

TP1 = S1



Motivation

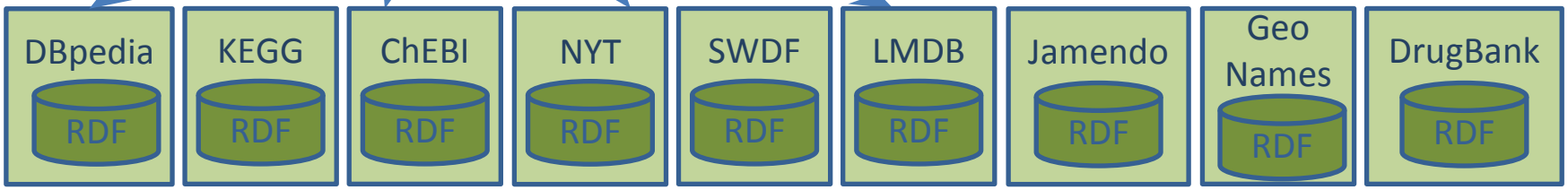
FedBench (LD3): Return for all US presidents their party membership and news pages about them.

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .      //TP1
?president dbpedia:nationality  dbpedia:United_States .  //TP2
?president dbpedia:party ?party .     //TP3
?x  nyt:topicPage ?page .                  //TP4
?x  owl:sameAs ?president .           //TP5
}
```

Source Selection Algorithm

Triple pattern-wise source selection

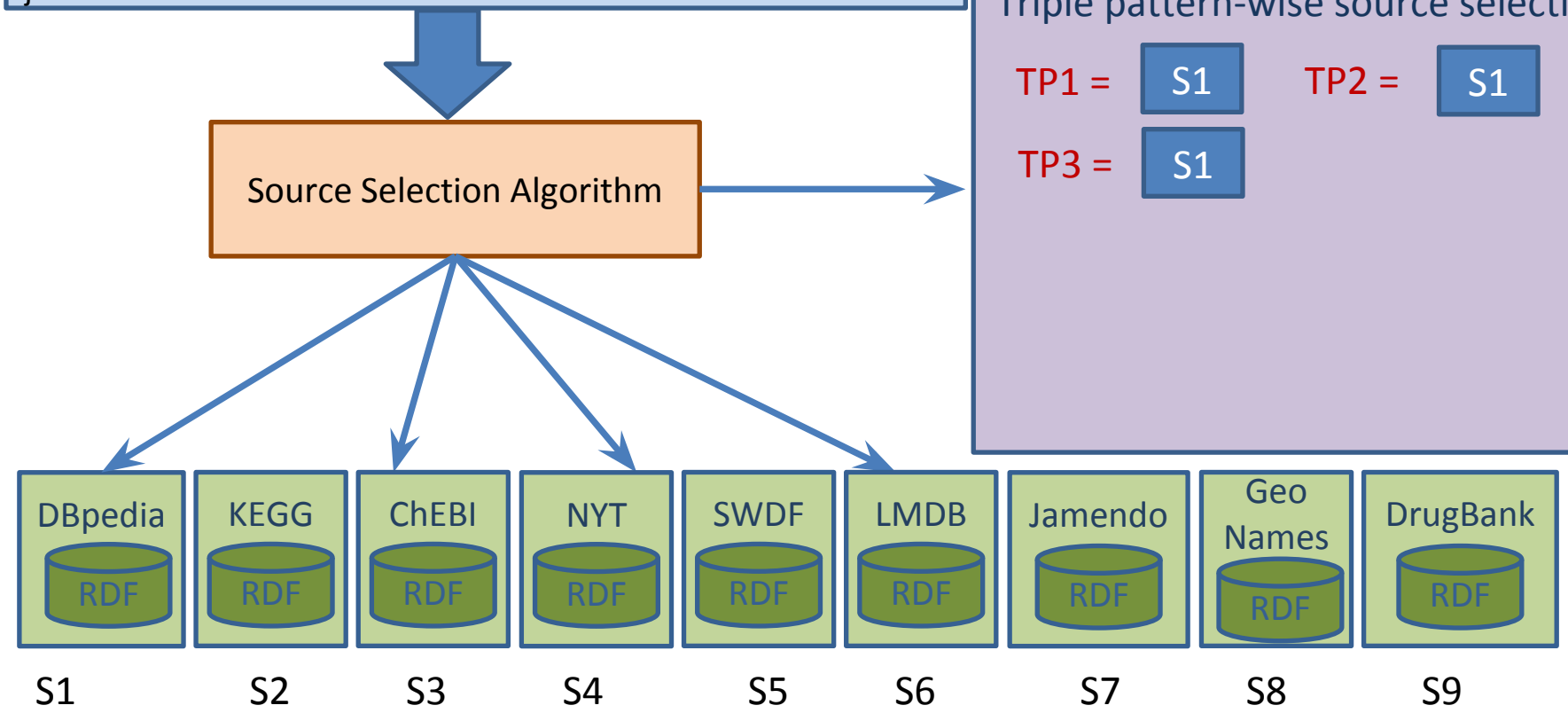
TP1 = S1 TP2 = S1



Motivation

FedBench (LD3): Return for all US presidents their party membership and news pages about them.

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .        //TP1
?president dbpedia:nationality  dbpedia:United_States . //TP2
?president dbpedia:party ?party .               //TP3
?x  nyt:topicPage ?page .                       //TP4
?x  owl:sameAs ?president .                     //TP5
}
```



Motivation

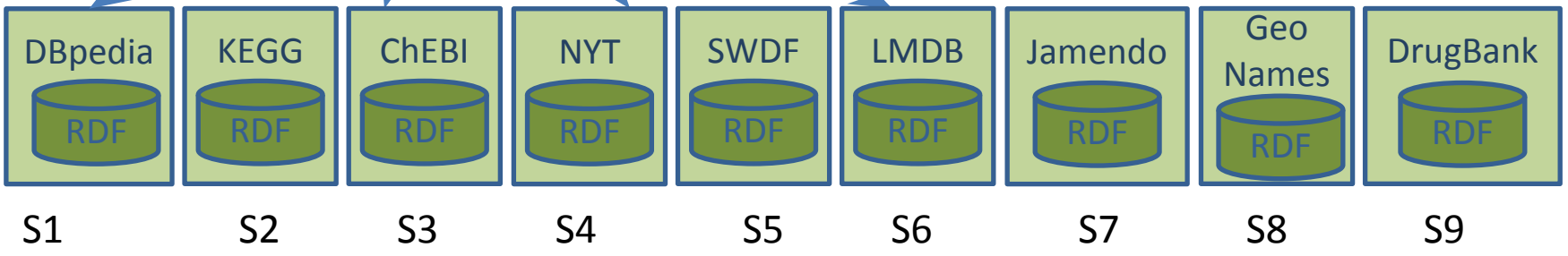
FedBench (LD3): Return for all US presidents their party membership and news pages about them.

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .        //TP1
?president dbpedia:nationality  dbpedia:United_States . //TP2
?president dbpedia:party ?party .     //TP3
?x  nyt:topicPage ?page .             //TP4
?x  owl:sameAs ?president .           //TP5
}
```

Source Selection Algorithm

Triple pattern-wise source selection

TP1 =	S1	TP2 =	S1
TP3 =	S1	TP4 =	S4



Motivation

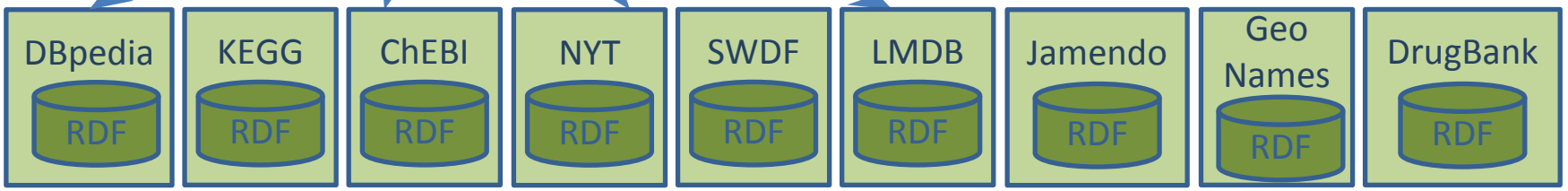
FedBench (LD3): Return for all US presidents their party membership and news pages about them.

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .        //TP1
?president dbpedia:nationality  dbpedia:United_States . //TP2
?president dbpedia:party ?party .     //TP3
?x  nyt:topicPage ?page .                 //TP4
?x  owl:sameAs ?president .             //TP5
}
```

Source Selection Algorithm

Triple pattern-wise source selection

TP1 =	S1	TP2 =	S1
TP3 =	S1	TP4 =	S4
TP5 =	S1	S2	S4	S5
	S6	S7	S8	S9



Motivation

FedBench (LD3): Return for all US presidents their party membership and news pages about them.

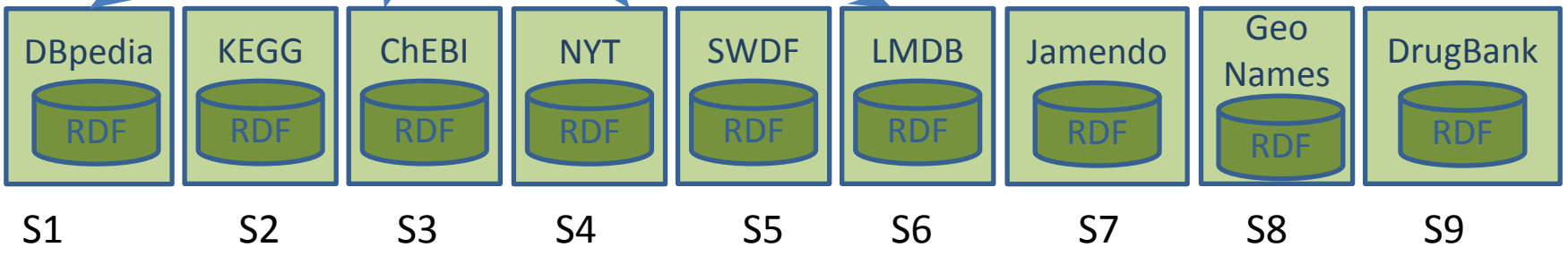
```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .        //TP1
?president dbpedia:nationality  dbpedia:United_States . //TP2
?president dbpedia:party ?party .     //TP3
?x  nyt:topicPage ?page .             //TP4
?x  owl:sameAs ?president .           //TP5
}
```

Source Selection Algorithm

Triple pattern-wise source selection

TP1 =	S1	TP2 =	S1
TP3 =	S1	TP4 =	S4
TP5 =	S1, S2, S4, S5, S6, S7, S8, S9		

Total selected sources: 12
Total SPARQL ASK queries: 9*5 = 45



Motivation

FedBench (LD3): Return for all US presidents their party membership and news pages about them.

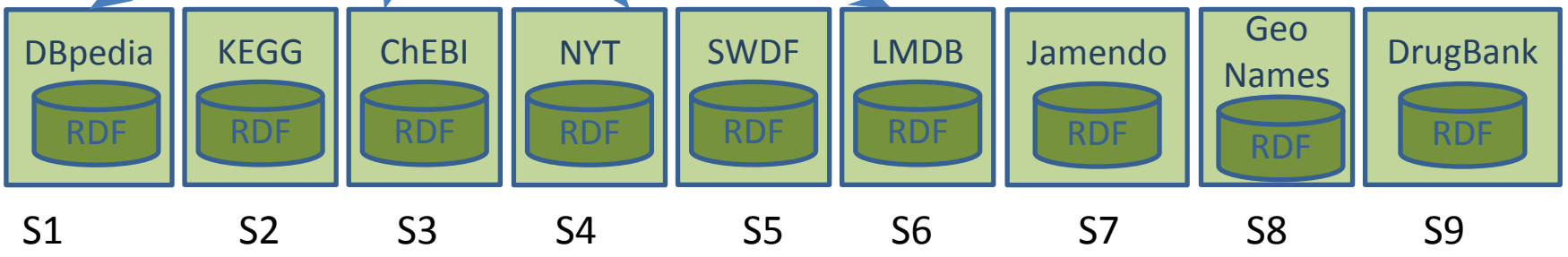
```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .        //TP1
?president dbpedia:nationality  dbpedia:United_States . //TP2
?president dbpedia:party ?party .     //TP3
?x  nyt:topicPage ?page .                 //TP4
?x  owl:sameAs ?president .             //TP5
}
```

Source Selection Algorithm

Triple pattern-wise source selection

TP1 =	S1	TP2 =	S1
TP3 =	S1	TP4 =	S4
TP5 =	S1	S2	S4	S5
	S6	S7	S8	S9

Total selected sources: 12
Total SPARQL ASK queries: 9*5 = 45



Motivation

FedBench (LD3): Return for all US presidents their party membership and news pages about them.

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .        //TP1
?president dbpedia:nationality  dbpedia:United_States . //TP2
?president dbpedia:party ?party .     //TP3
?x  nyt:topicPage ?page .                 //TP4
?x  owl:sameAs ?president .             //TP5
}
```

Source Selection Algorithm

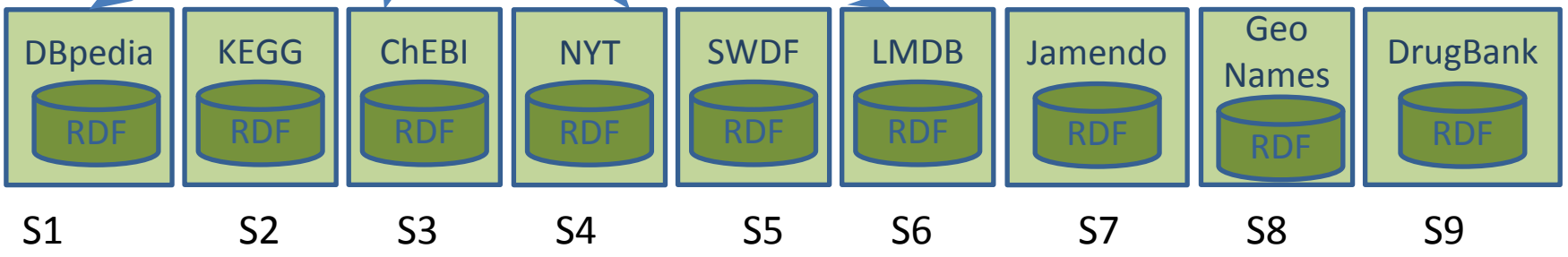
Triple pattern-wise source selection

TP1 = S1 TP2 = S1

TP3 = S1 TP4 = S4

TP5 = S4

Optimal selected sources: 5



Problem Statement

- An overestimation of the number of contributing sources can be expensive
  - Resources are wasted
  - Query runtime is increased
  - Extra traffic is generated
- How can we perform a join-aware triple-pattern-wise source selection in a time-efficient way?

Key Concept

- Make use URI authority

http://dbpedia.org/ontology/party

Scheme Authority Path

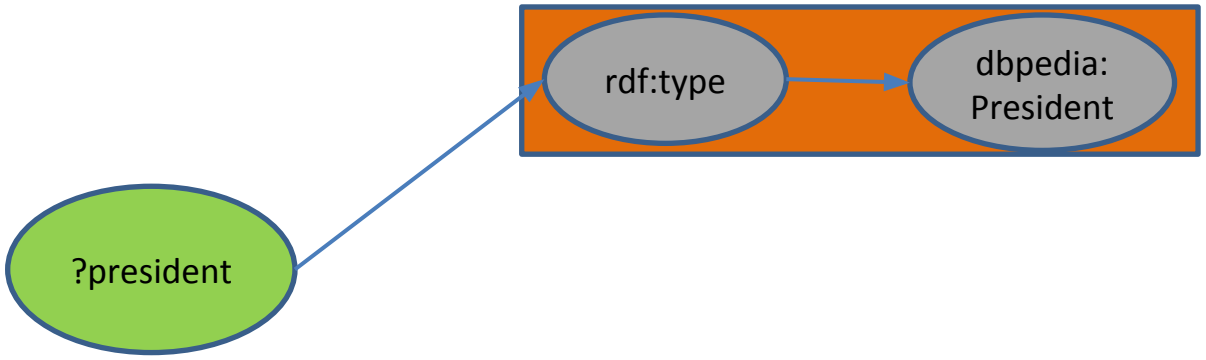
For URI details: http://tools.ietf.org/html/rfc3986

Data Summaries

```
[] a ds:Service ;
ds:endpointUrl <http://dbpedia.org/sparql> ;
ds:capability [   ds:predicate  dbpedia:party ;
                  ds:sbjAuthority  <http://dbpedia.org/> ;
                  ds:objAuthority  <http://dbpedia.org/> ;  ] ;
ds:capability [   ds:predicate rdf:type ;
                  ds:sbjAuthority  <http://dbpedia.org/> ;
                  ds:objAuthority  owl:Thing, dbpedia:President; ] ;
#all distinct classes
ds:capability [   ds:predicate  dbpedia:postalCode ;
                  ds:sbjAuthority  <http://dbpedia.org/> ;  ] ;
#No objAuthority as the object value for dbpedia:postalCode is string
```

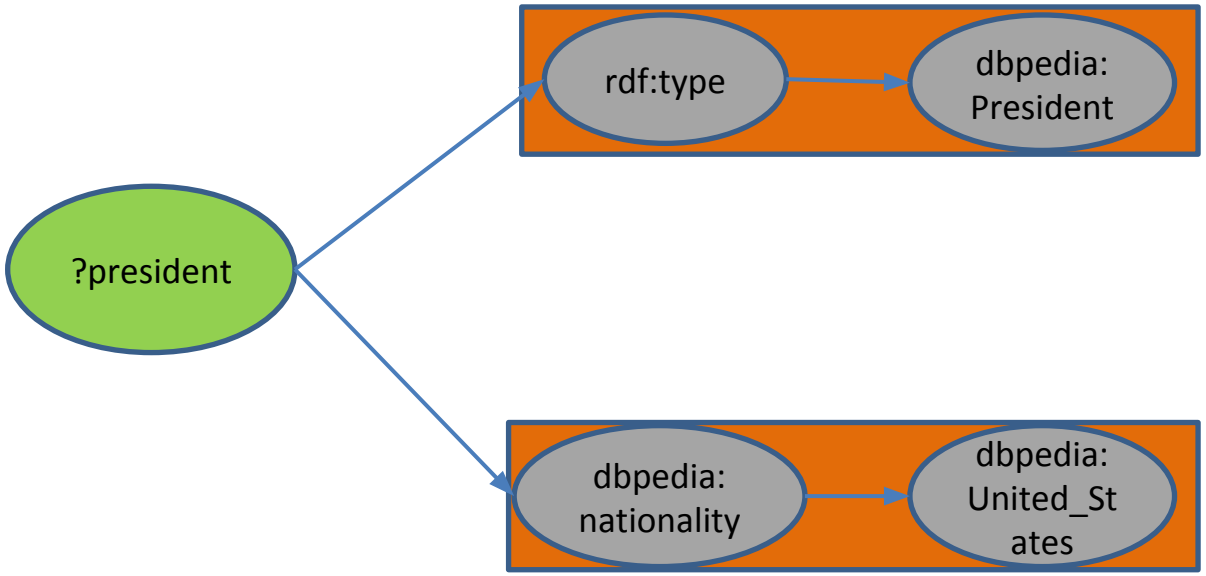
HiBISCuS: SPARQL Query as Hypergraph

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



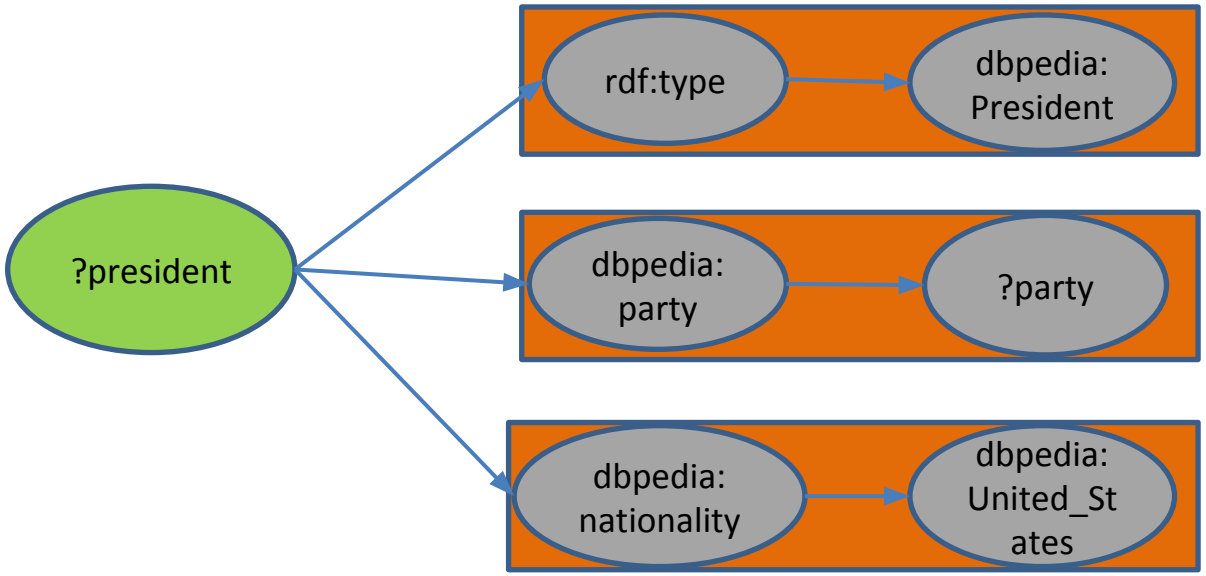
HiBISCuS: SPARQL Query as Hypergraph

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



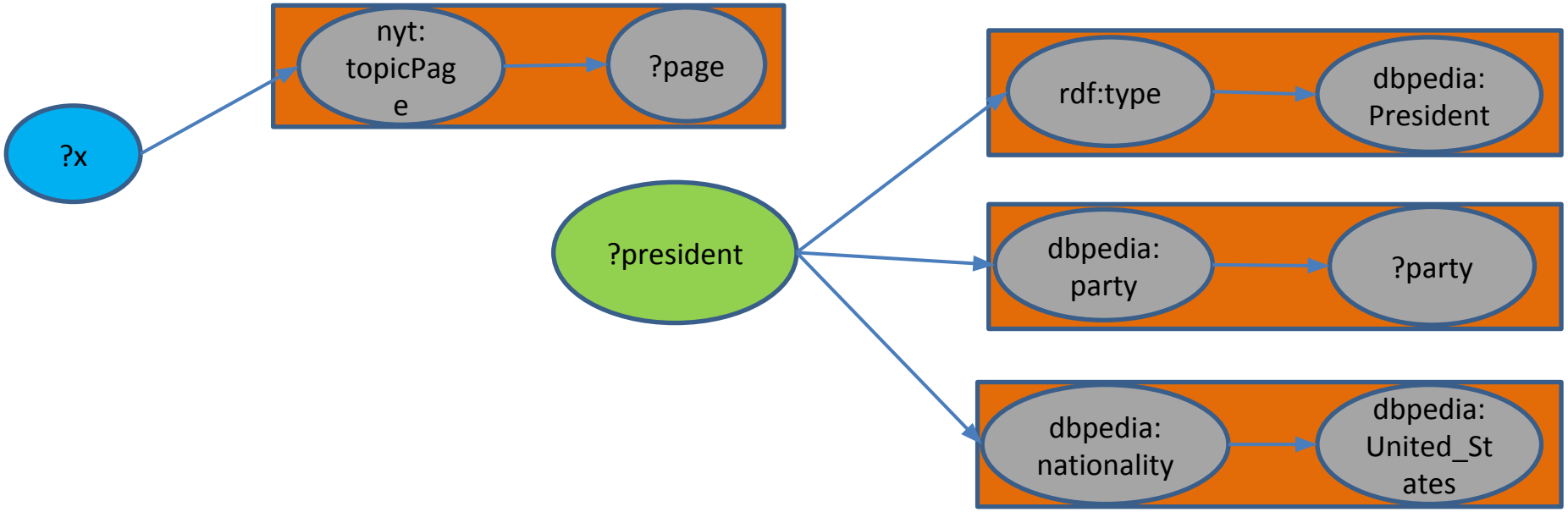
HiBISCuS: SPARQL Query as Hypergraph

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



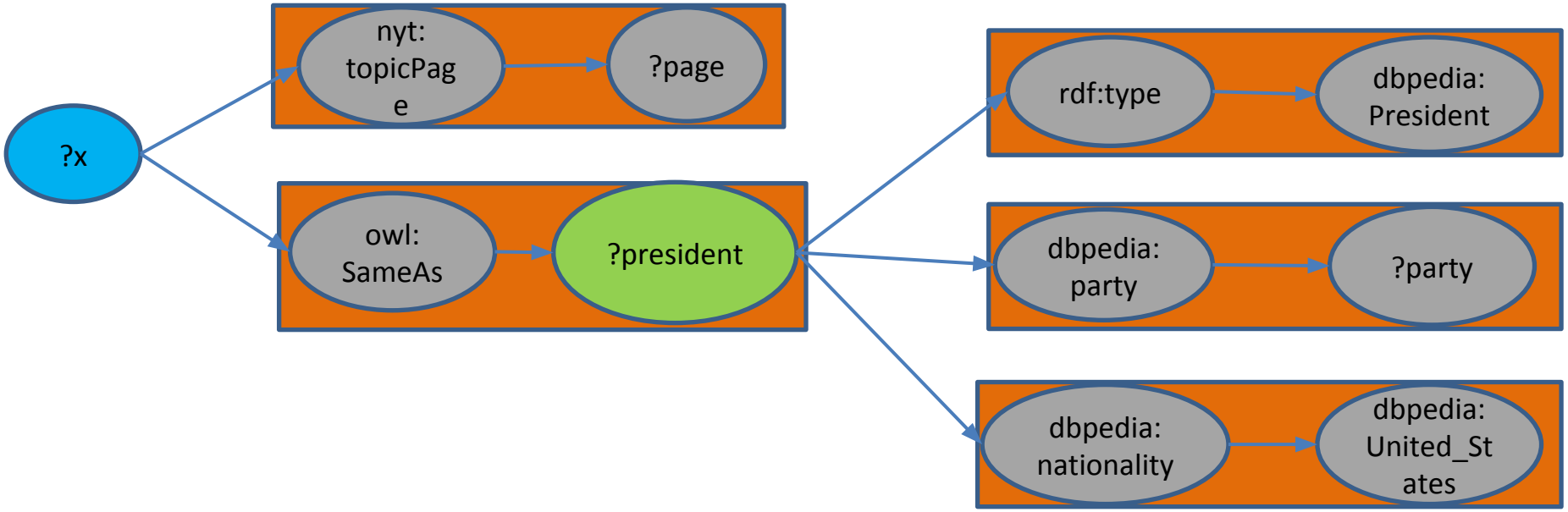
HiBISCuS: SPARQL Query as Hypergraph

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



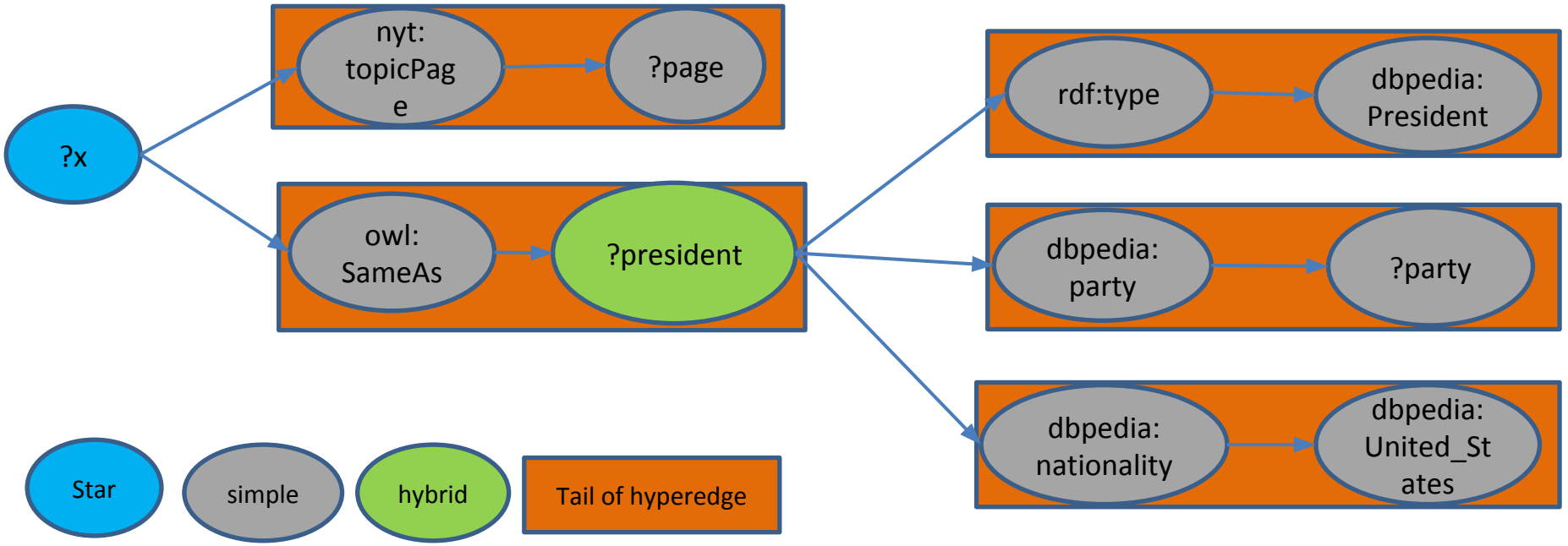
HiBISCuS: SPARQL Query as Hypergraph

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



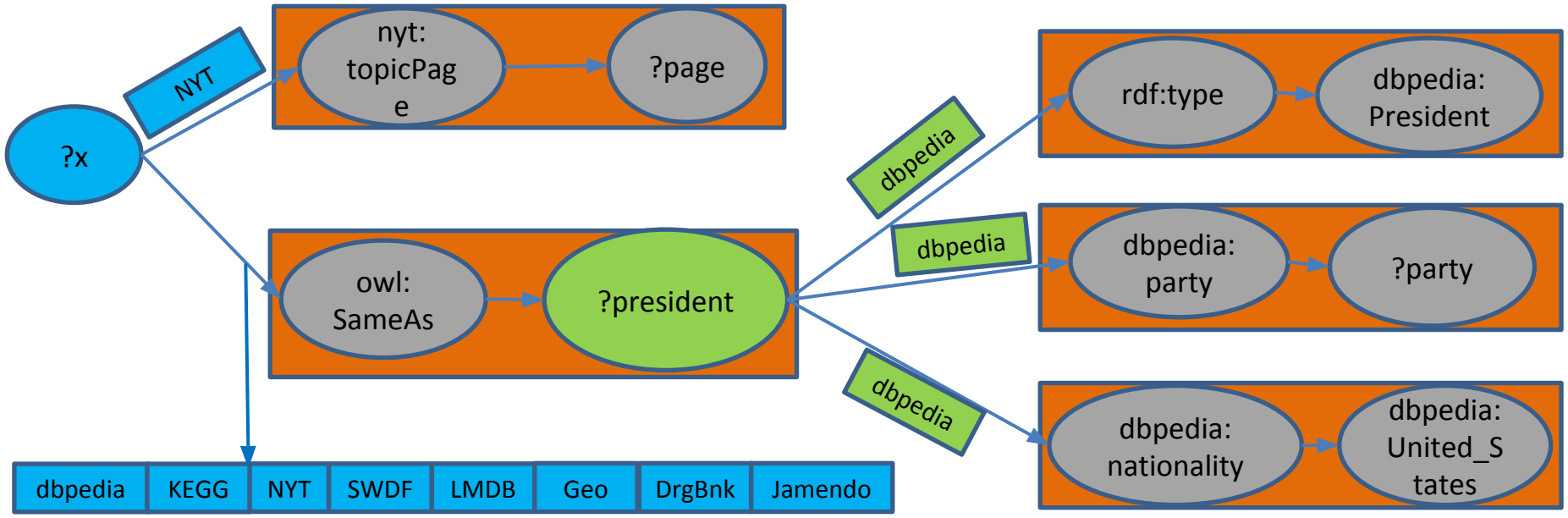
HiBISCuS: SPARQL Query as Hypergraph

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



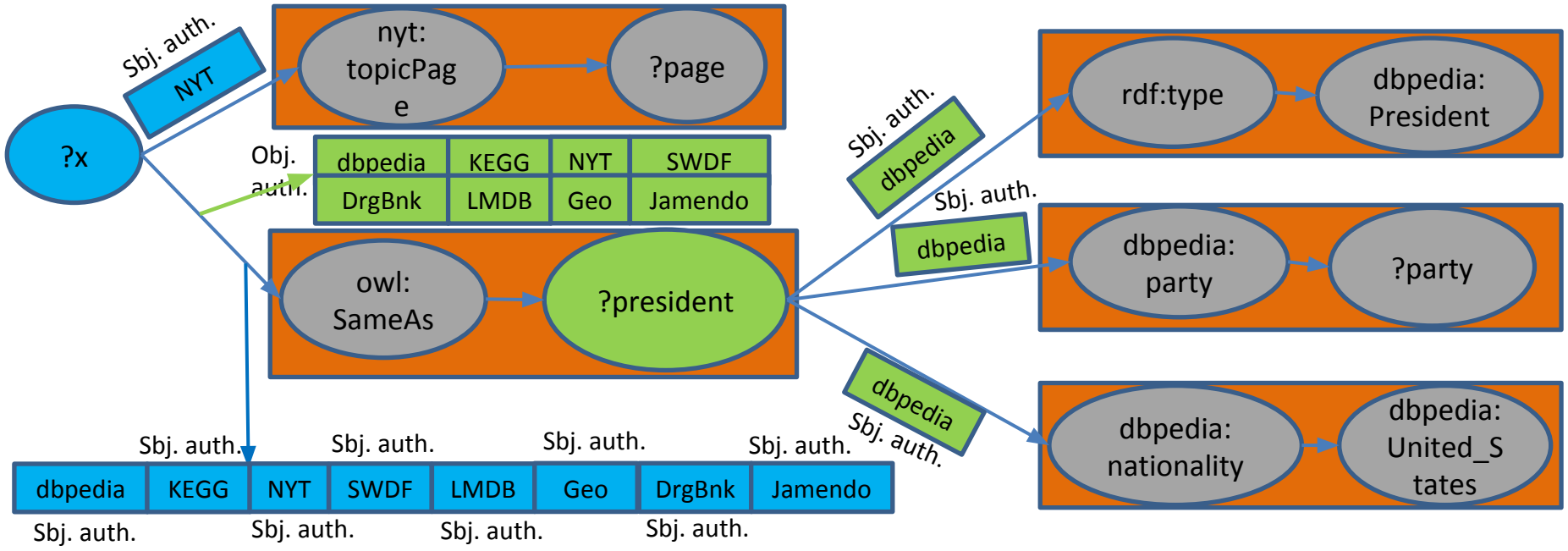
HiBISCuS: Source Selection

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



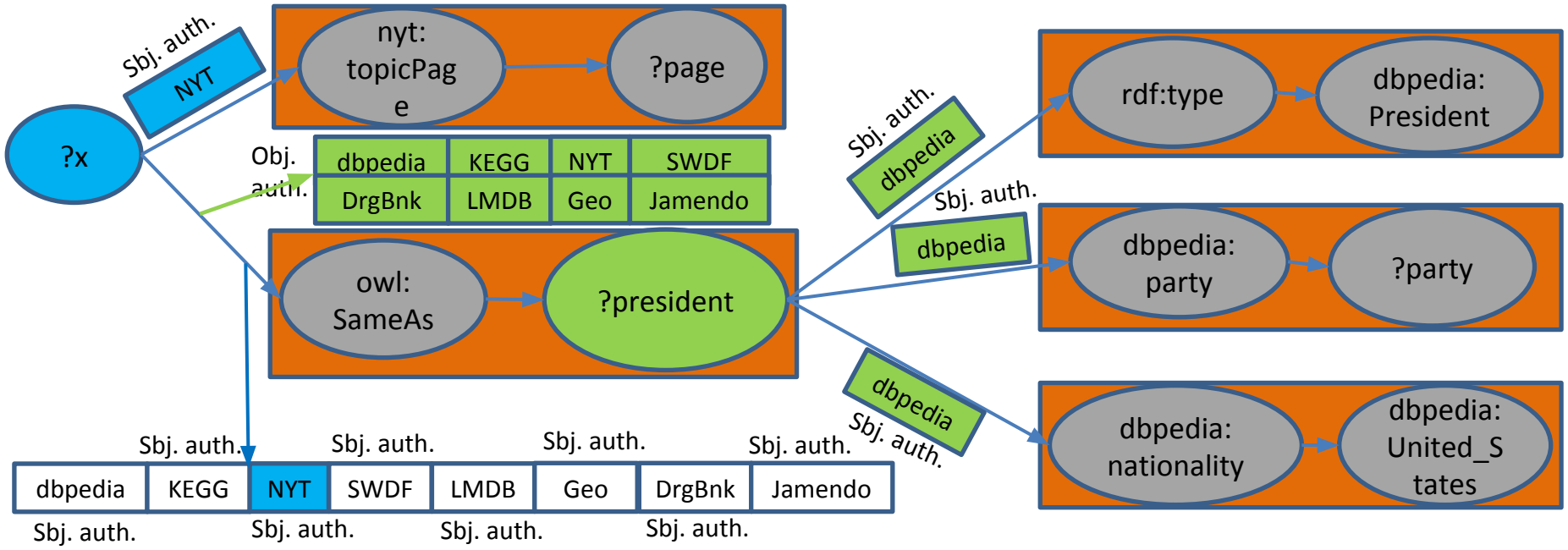
HiBISCuS: Source Pruning

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



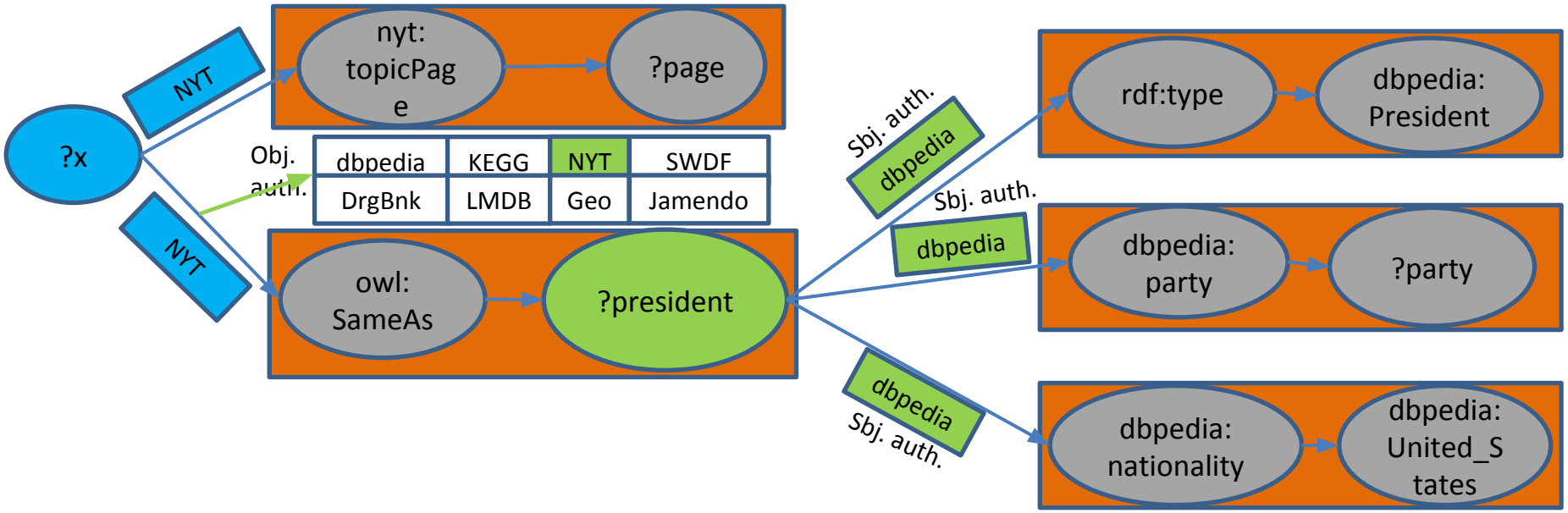
HiBISCuS: Source Pruning

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



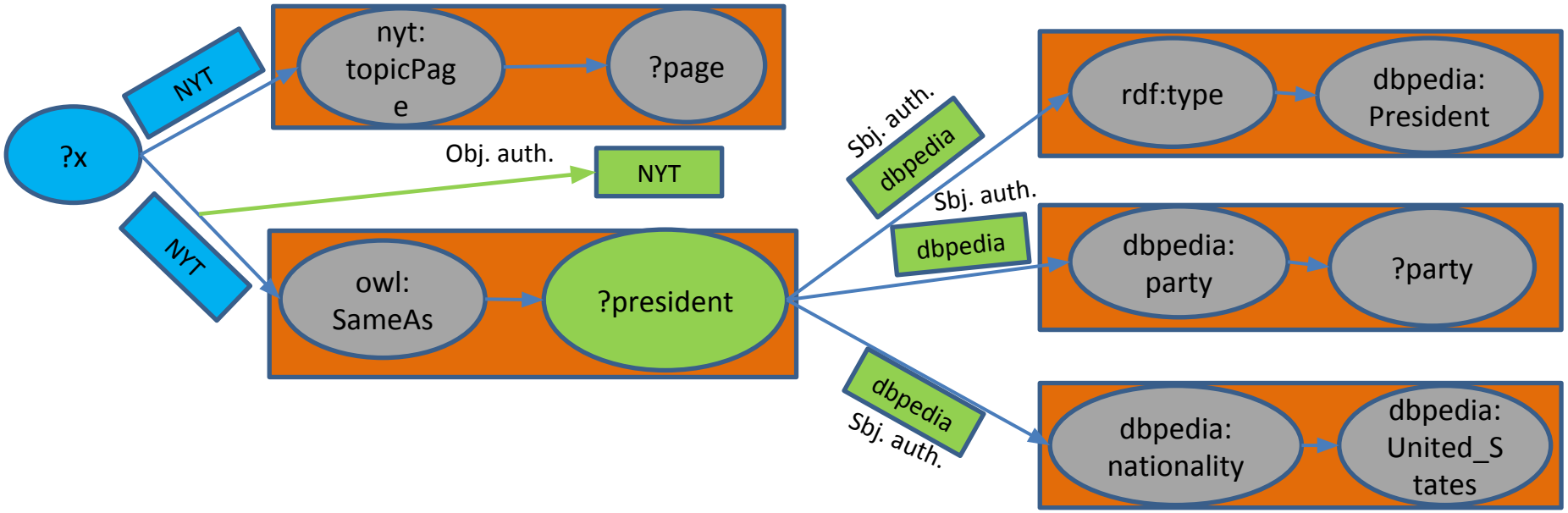
HiBISCuS: Source Pruning

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



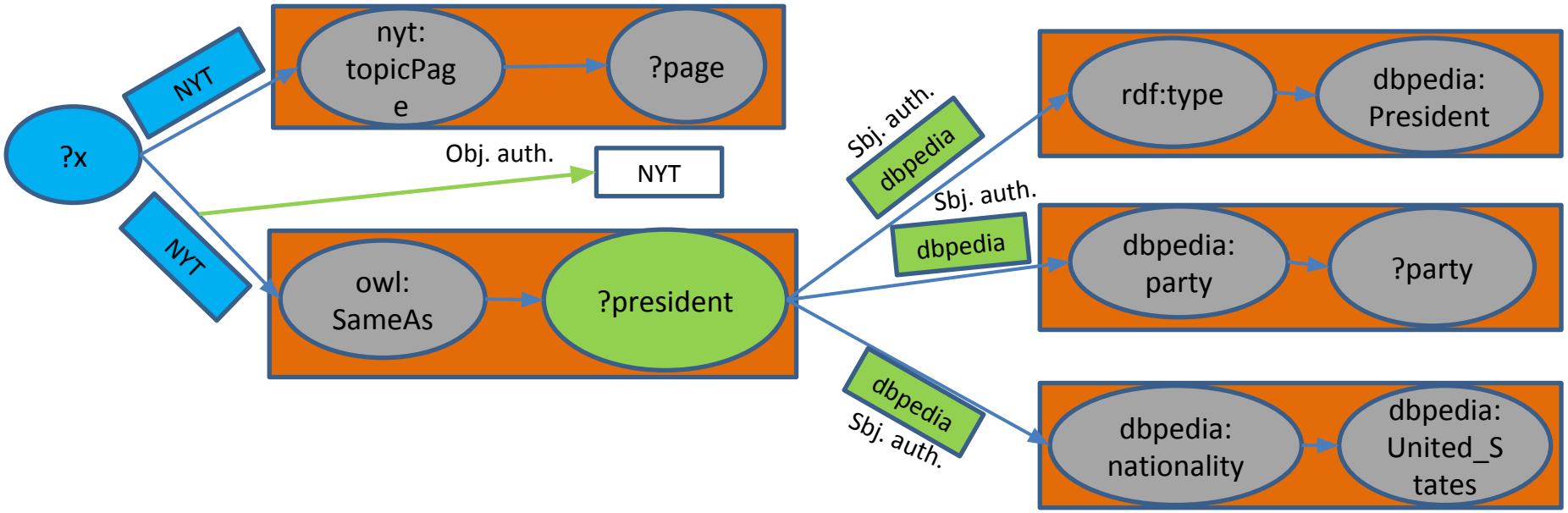
HiBISCuS: Source Pruning

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



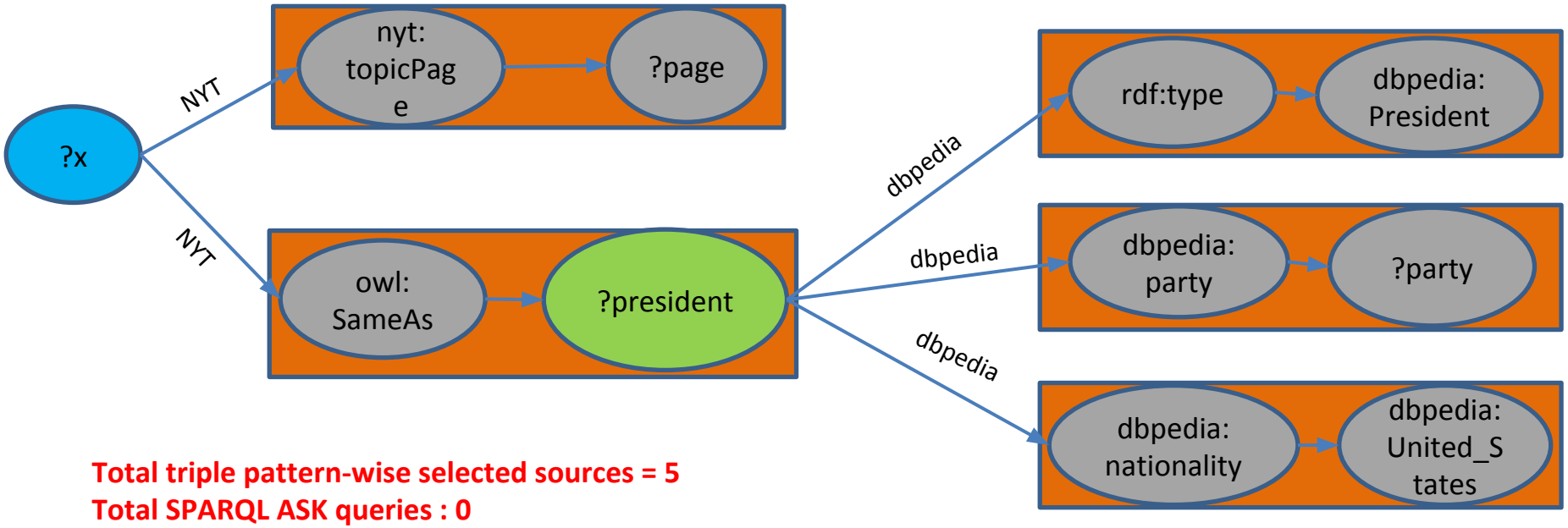
HiBISCuS: Source Pruning

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



HiBISCuS: Source Pruning

```
SELECT ?president ?party ?page
WHERE {
?president  rdf:type dbpedia:President .
?president dbpedia:nationality  dbpedia:United_States .
?president dbpedia:party ?party .
?x  nyt:topicPage ?page .
?x  owl:sameAs ?president .
}
```



Total triple pattern-wise selected sources = 5
Total SPARQL ASK queries : 0

Experimental Setup

- Benchmark
  - FedBench
  - Real-world datasets collection
  - Real queries showing typical request
  - Used all of the 25 queries
- HiBISCuS Extensions
  - FedX 2.0
  - DARQ
  - SPLENDID
- We also included ANAPSID (version of 12/2013) without HiBISCus extension

Experimental Setup

- Metrics
  - Index generation time
  - Index size
  - Total triple pattern-wise sources selected
  - Total number of SPARQL ASK requests used
  - Source selection time
  - Query execution time

Index Generation Time and Compression Ratio

	FedX	SPLENDID	LHD	DARQ	ANAPSID	ADERIS	Qtree	HiBISCus
Index Generation Time (min)	NA	75	75	102	6	6	-	36
Compression Ratio	NA	99.998	99.998	99.997	99.999	99.999	96.000	99.997

Compression Ratio = 1 - index size/ total datadump size

Efficient Source Selection

FedBench Cross Domain Queries			
System	#TP	#AR	SST(ms)
FedX(cold)	78	252	317.8
FedX(warm)	78	0	7.3
SPLENDID	78	99	320.8
DARQ	84	0	7.2
ANAPSID	36	43	186
HiBISCuS(cold)	35	27	63
HiBISCuS(warm)	35	0	30.4

FedBench Linked Data Queries			
System	#TP	#AR	SST(ms)
FedX(cold)	97	369	307.8
FedX(warm)	97	0	8
SPLENDID	97	126	279
DARQ	113	0	7.7
ANAPSID	54	37	803.5
HiBISCuS(cold)	47	9	20.6
HiBISCuS(warm)	47	0	16

FedBench Life Sciences Queries			
System	#TP	#AR	SST(ms)
FedX(cold)	56	297	375.5
FedX(warm)	56	0	8.2
SPLENDID	56	90	307.2
DARQ	77	0	7.5
ANAPSID	44	63	477.4
HiBISCuS(cold)	41	18	31.8
HiBISCuS(warm)	41	0	23.1

Overall			
System	#TP	#AR	SST(ms)
FedX(cold)	231	918	330
FedX(warm)	231	0	8
SPLENDID	231	315	299
DARQ	274	0	7.5
ANAPSID	134	143	554
HiBISCuS(cold)	123	54	35.6
HiBISCuS(warm)	123	0	22

Efficiency of Source Selection

FedBench Cross Domain Queries

System	#TP	#AR	SST(ms)
FedX(cold)	78	252	317.8
FedX(warm)	78	0	7.3
SPLENDID	78	99	320.8
DARQ	84	0	7.2
ANAPSID	36	43	186
HiBISCuS(cold)	35	27	63
HiBISCuS(warm)	35	0	30.4

FedBench Linked Data Queries

System	#TP	#AR	SST(ms)
FedX(cold)	97	369	307.8
FedX(warm)	97	0	8
SPLENDID	97	126	279
DARQ	113	0	7.7
ANAPSID	54	37	803.5
HiBISCuS(cold)	47	9	20.6
HiBISCuS(warm)	47	0	16

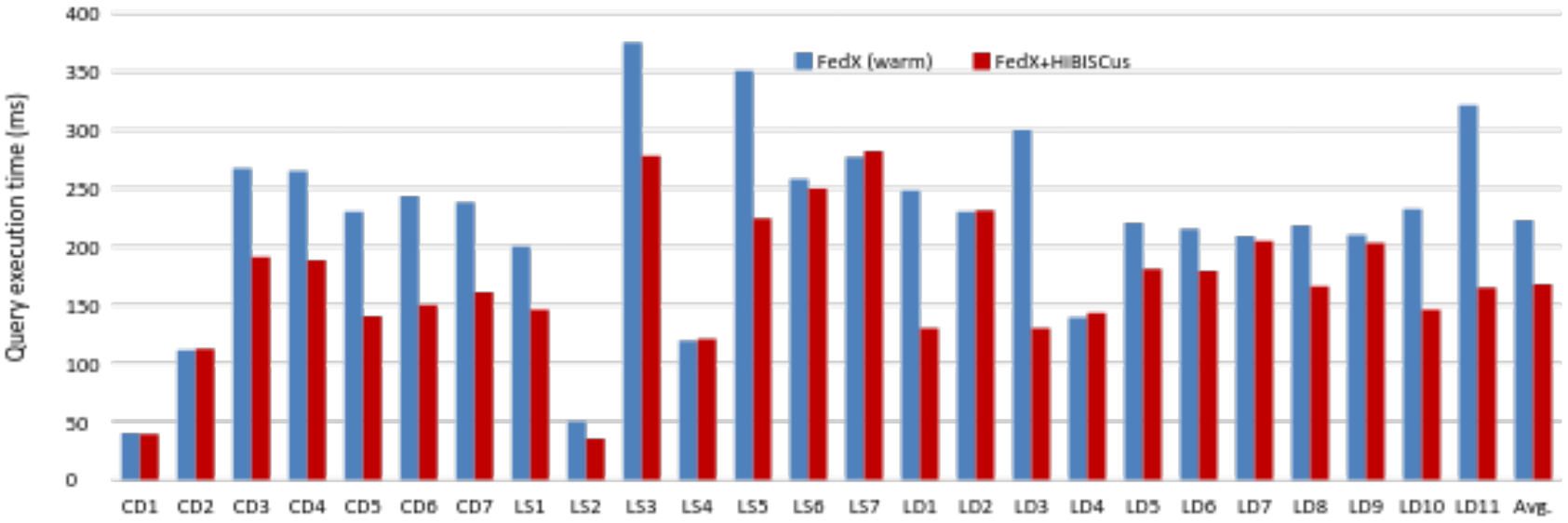
FedBench Life Sciences Queries

System	#TP	#AR	SST(ms)
FedX(cold)	56	297	375.5
FedX(warm)	56	0	8.2
SPLENDID	56	90	307.2
DARQ	77	0	7.5
ANAPSID	44	63	477.4
HiBISCuS(cold)	41	18	31.8
HiBISCuS(warm)	41	0	23.1

Overall

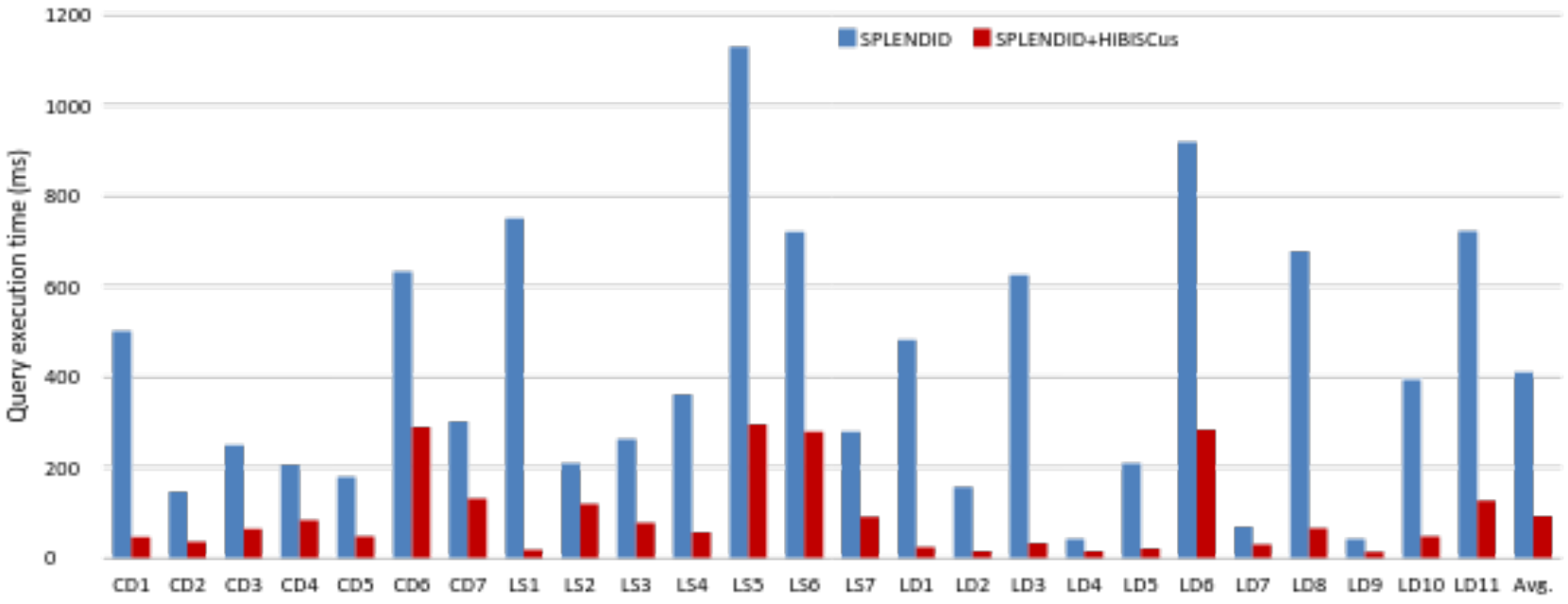
System	#TP	#AR	SST(ms)
FedX(cold)	231	918	330
FedX(warm)	231	0	8
SPLENDID	231	315	299
DARQ	274	0	7.5
ANAPSID	134	143	554
HiBISCuS(cold)	123	54	35.6
HiBISCuS(warm)	123	0	22

FedX Extension with HiBISCuS



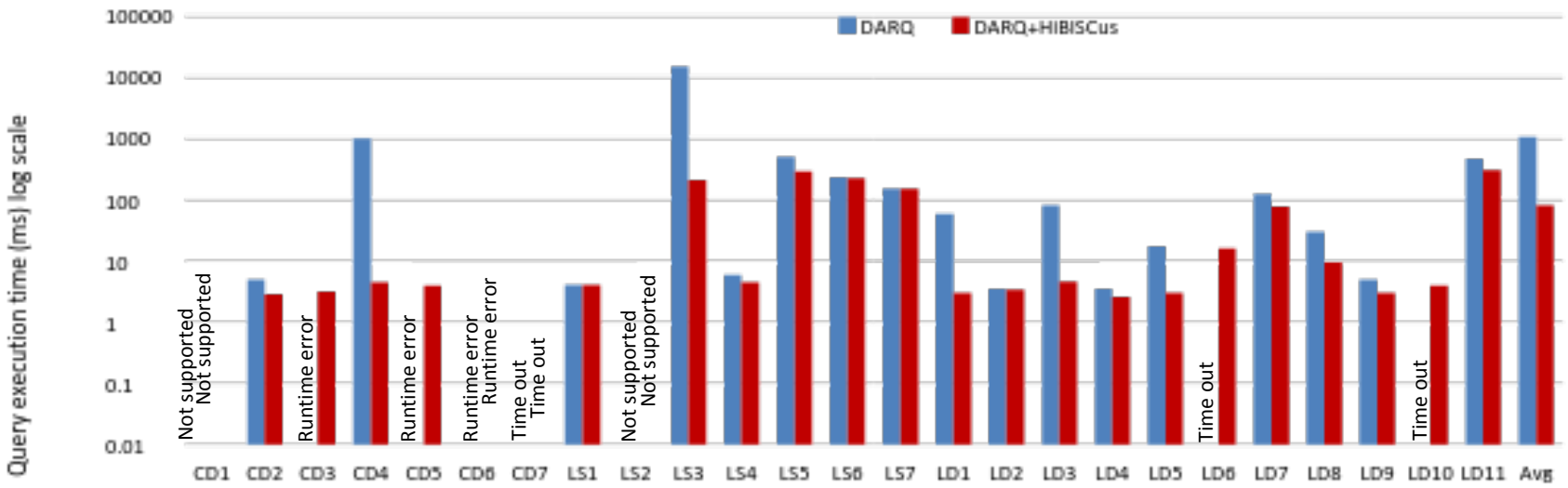
Improvement in 20/25 queries with net performance improvement 24.61%

SPLENDID Extension with HiBISCuS



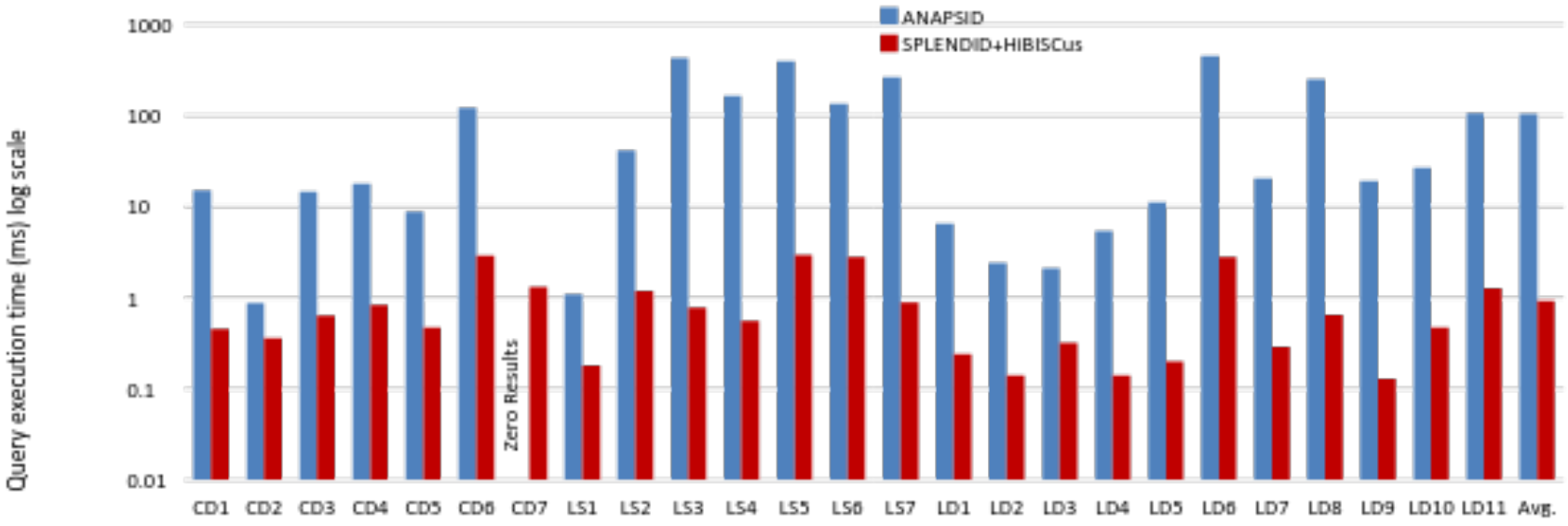
Improvement in 25/25 queries with net performance improvement 82.72%

DARQ Extension with HiBISCuS



Improvement in 21/21 queries with net performance improvement 92.22%

SPLENDID+HiBISCuS vs ANAPSID



Improvement in 24/24 queries with net performance improvement 98%

Conclusion and Future Work

- An overestimation of triple pattern-wise source selection can be expensive
- Join-aware triple pattern-wise source selection is more efficient than simple triple pattern-wise source selection
- Performance improvements
  - FedX : 24.61 %
  - SPLENDID: 82.72 %
  - DARQ: 92.22 %
  - SPLENDID+HiBISCus is 98% faster than ANAPSID
- BigRDFBench is on the roadmap

Thank you!
Questions?

Axel Ngonga
AKSW Research Group
University of Leipzig
ngonga@informatik.uni-leipzig.de
http://aksw.org
https://code.google.com/p/hibiscusfederation/

Source Selection

- Triple pattern-wise source selection
  - Ensures 100% recall
  - Can over-estimate capable sources
  - Can be expensive, e.g., total number of SPARQL ASK requests used
  - Performed by FedX, SPLENDID, LHD, DARQ, ADERIS etc.
- Join-aware triple-pattern wise source selection
  - Ensures 100% recall
  - May selects optimal/close to optimal capable sources
  - Can be expensive, e.g., total number of SPARQL ASK requests used
  - Can significantly reduce the query execution time
  - Performed by ANAPSID