Geographica: A Benchmark for Geospatial RDF Stores

George Garbis, Kostis Kyzirakos, Manolis Koubarakis

Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece

12th International Semantic Web Conference
(Evaluation Track)

Outline

- **Motivation**
- The benchmark Geographica
  - Real-world workload
  - Synthetic workload
- Evaluating the performance of geospatial RDF stores using Geographica
- Conclusions

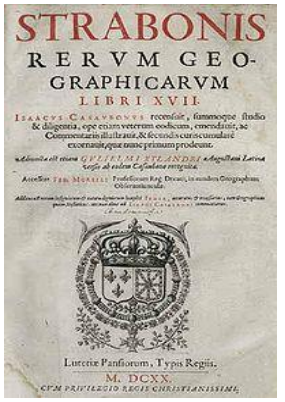
Motivation

- Lots of **geospatial data is available on the Web** today.
- Lots of geospatial data is quickly being **transformed into linked geospatial data!**
- People have started building **applications** using such data.
- Geospatial extensions of SPARQL (e.g., **GeoSPARQL** and **stSPARQL**) have been recently developed.
- RDF stores provide support for GeoSPARQL (e.g., **Strabon**, **Oracle 12c, uSeekM**, **Parliament**) or provide limited geospatial functionality (e.g., Virtuoso, BigOwlim, AllegroGraph)

The Benchmark Geographica

- Aim: measure the performance of **today's geospatial RDF stores**
- Organized around two workloads:
  - **Real-world** workload:
    - Based on existing linked geospatial datasets and known application scenarios
  - **Synthetic** workload:
    - Measure performance in a controlled environment where we can play around with selectivity of queries.

- **Γεωγραφικά:** 17-volume geographical encyclopedia by Στράβων (AD 17)



Outline

- Motivation
- The benchmark Geographica
  - **Real-world workload**
  - Synthetic workload
- Evaluating the performance of geospatial RDF stores using Geographica
- Conclusions

Real-World Workload
Datasets

- **Datasets:** Real-world datasets for the geographic area of Greece playing an **important role in the LOD** cloud or **having complex geometries**
  - LinkedGeoData (**LGD**) for rivers and main roads in Greece
  - **GeoNames** for Greece
  - **DBpedia** for Greece
  - Greek Administrative Geography (**GAG**)
  - CORINE land cover (**CLC**) for Greece
  - Hotspots

Real-World Workload
Datasets

Dataset	Size	# of Triples	# of Points	# of Lines (max/min/avg points/line)	# of Polygons (max/min/avg points/polygon)
GeoNames	45MB	400K	22K	-	-
Dbpedia	89MB	430K	8K	-	-
LGD	29MB	150K	-	12K (1.6K/2/21)	-
GAG	33MB	4K	-	-	325 (15K/4/400)
CLC	401MB	630K	-	-	45K (5K/4/140)
Hotspots	90MB	450K	-	-	37K (4/4/4)

Real-World Workload
Parts

- For this workload, Geographica has two parts (following Jackpine):
  - **Micro part:** Tests primitive spatial functions offered by geospatial RDF stores
  - **Macro part:** Simulates some typical application scenarios

Real-World Workload
Micro part

- **29 SPARQL queries** that consist of **one or two triple patterns and a spatial function.**
- Functions included:
  - **Non-topological:** `boundary, envelope, convex hull, buffer, area`
  - **Topological:** `equals, intersects, overlaps, crosses, within, distance, disjoint`
  - **Spatial aggregates:** `extent, union`
- These functions are used for **spatial selections** and **spatial joins**

Example – non-topological
Micro part

- Construct the boundary of all polygons of CLC

```
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
PREFIX clc: <http://geo.linkedopendata.gr/corine/ontology#>

SELECT ( geof:boundary(?o1)  as ?ret )
WHERE {
      GRAPH dataset:clc  { ?s1 clc:asWKT ?o1. }
}
```

Example – spatial selection
Micro part

- Find all points in GeoNames that are within a given polygon.

```
PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>

PREFIX geonames: <http://www.geonames.org/ontology#>


SELECT ?s1 ?o1

WHERE {

   GRAPH dataset:geonames { ?s1 geonames:asWKT ?o1 }

   FILTER( geof:sfWithin(?o1, "POLYGON((…))"^^geo:wktLiteral)).

}
```

Example – spatial join
Micro part

- Find all pairs of GAG polygons that overlap

```
PREFIX dataset: <http://geographica.di.uoa.gr/dataset/>
PREFIX gag: <http://geo.linkedopendata.gr/gag/ontology/>
PREFIX clc: <http://geo.linkedopendata.gr/corine/ontology#>
SELECT ?s1 ?s2
WHERE {
   GRAPH dataset:gag {?s1 gag:asWKT ?o1}
   GRAPH dataset:clc   {?s2 clc:asWKT ?o2}
   FILTER( geof:sfOverlaps(?o1, ?o2) )
}
```

Real-World Workload
Micro part

- **Spatial Selections**

	Query Point	Query Line	Query Polygon
Points	Within Buffer Distance		Within Disjoint
Lines		Equals Crosses	Intersects Disjoint
Polygons		Intersects	Equals Overlaps

- **Spatial Joins**

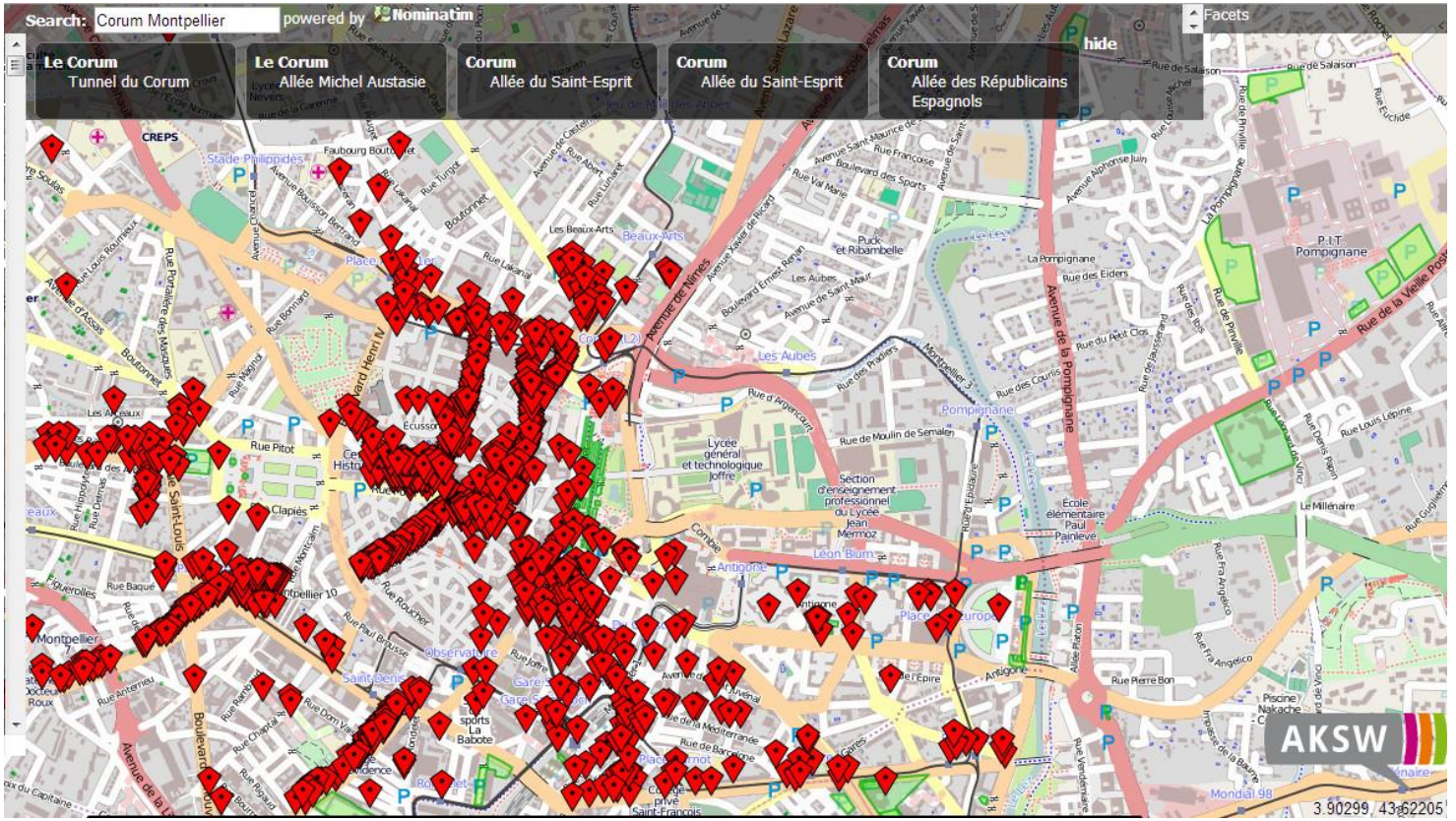
	Points	Lines	Polygons
Points	Equals	Intersects	Intersects Within
Lines			Intersects Within Crosses
Polygons			Within Touches Overlaps

Real-World Workload
Macro part: Scenarios

- **Reverse Geocoding**

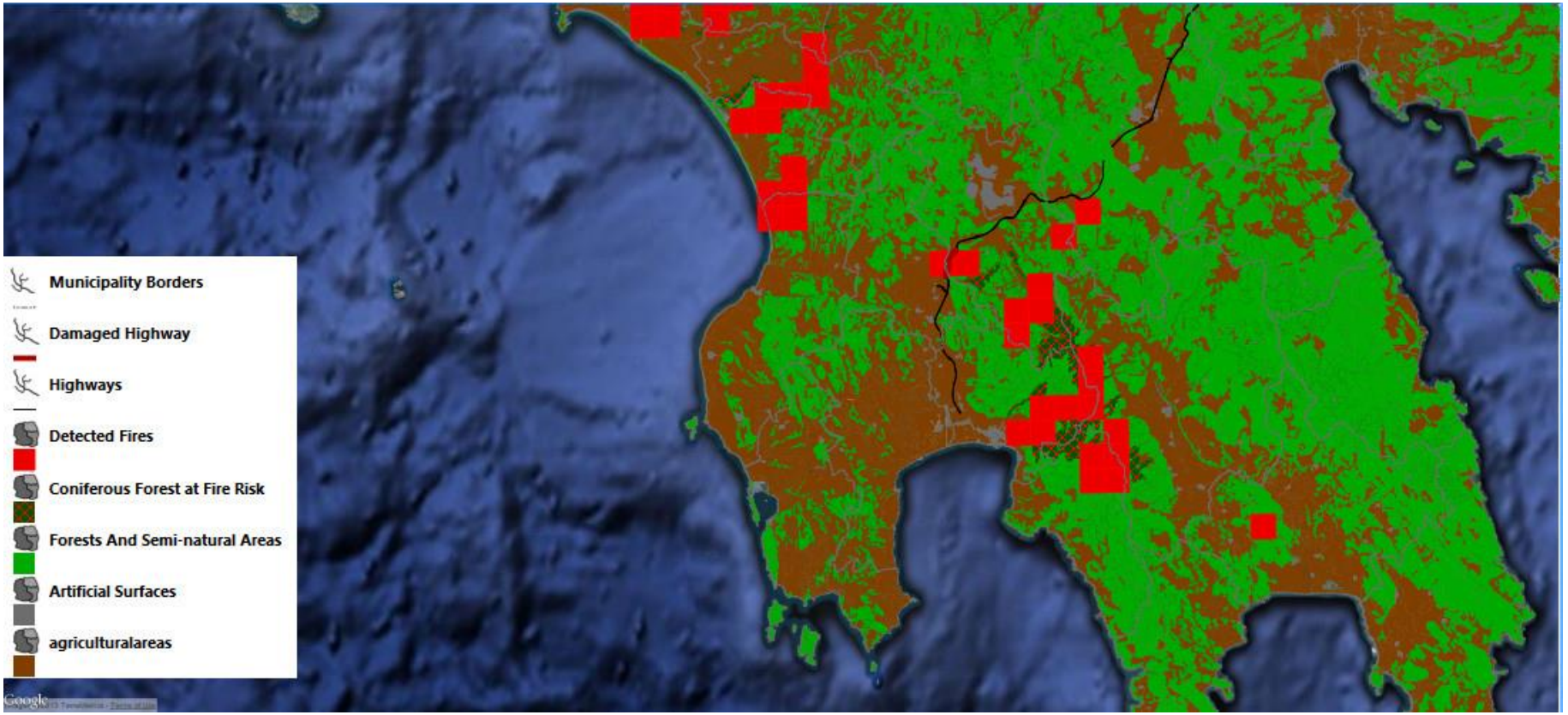
Real-World Workload
Macro part: Scenarios

- **Reverse Geocoding**
- **Web Map Search and Browsing**



Real-World Workload
Macro part: Scenarios

- **Reverse Geocoding**
- **Web Map Search and Browsing**
- **Rapid Mapping for Fire Monitoring**



Outline

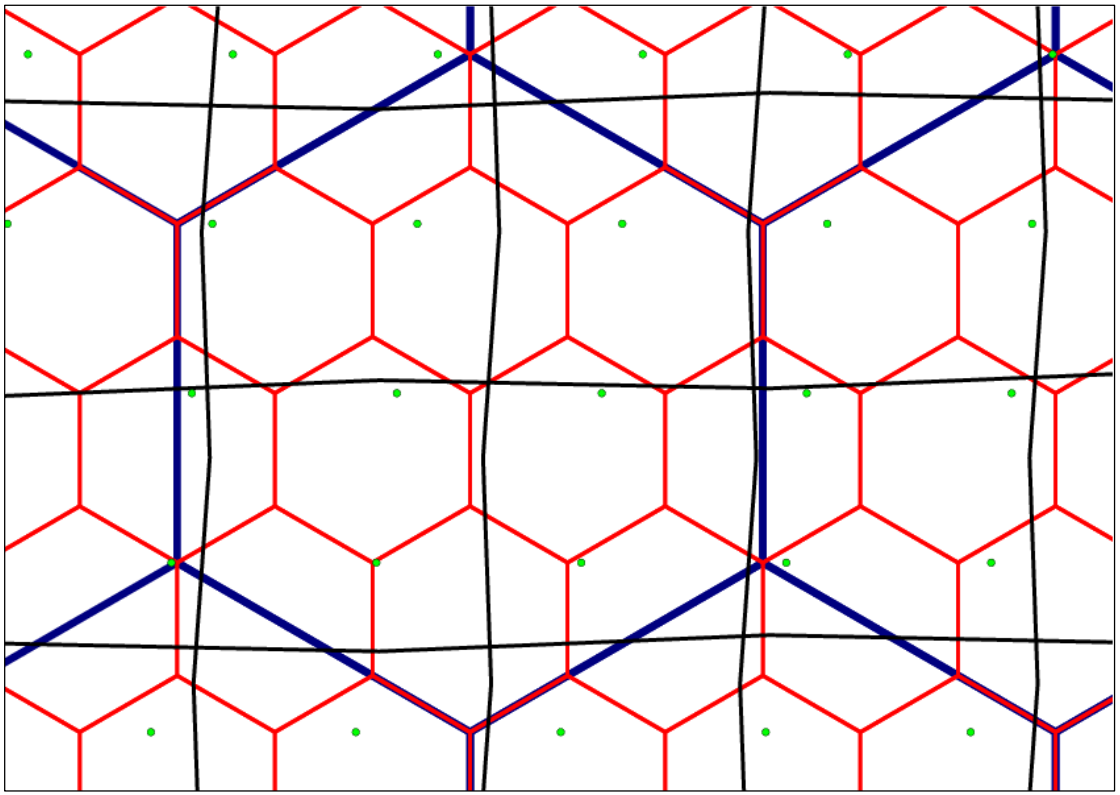
- Motivation
- The benchmark Geographica
  - Real-world workload
  - **Synthetic workload**
- Evaluating the performance of geospatial RDF stores using Geographica
- Conclusions

Synthetic Workload

- **Goal:** Evaluate performance in a controlled environment where we can vary the thematic and spatial selectivity of queries
  - **Thematic selectivity:** the fraction of the total geographic features of a dataset that satisfy the non-spatial part of a query
  - **Spatial selectivity:** the fraction of the total geographic features of a dataset which satisfy the topological relation in the FILTER clause of a query

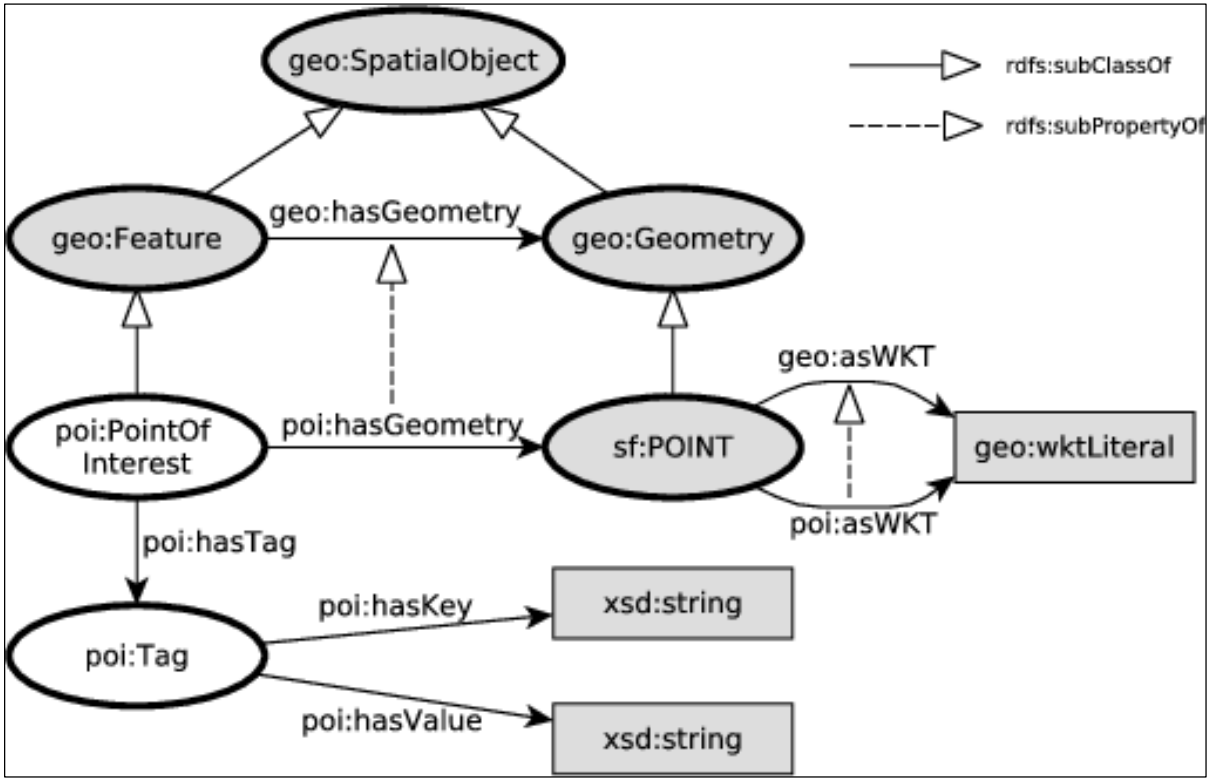
Synthetic Workload
Generator

- **Dataset**: As in VESPA, the produced datasets are geographic features on a synthetic map:
  - States in a country ($(n/3)^2$)
  - Land ownership (n^2)
  - Roads (n)
  - POI (n^2)



Synthetic Workload
Ontology

- Based roughly on the ontology of OpenStreetMap and the GeoSPARQL vocabulary
- Tagging each feature with a key enables us to select a known fraction of features in a uniform way



Synthetic Workload
Query template for spatial selections

```
SELECT ?s

WHERE {
   ?s ns:hasGeometry ?g.
   ?s c:hasTag ?tag.
   ?g ns:asWKT ?wkt.
   ?tag ns:hasKey "THEMA"

FILTER(FUNCTION(?wkt, "GEOM"^^geo:wktLiteral))}
```

- Parameters:
  - **ns:** specifies the kind of feature (and geometry type) examined
  - **FUNCTION:** specifies the topological function examined
  - **THEMA:** defines the **thematic selectivity** of the query using another parameter **k**
  - **GEOM:** specifies a rectangle that controls the **spatial selectivity** of the query

Synthetic Workload
Query template for spatial joins

```
SELECT ?s1 ?s2

WHERE {
  ?s1 ns1:hasGeometry ?g1.
  ?s1 ns1:hasTag ?tag1.
  ?g1 ns1:asWKT ?wkt1.
  ?tag1 ns1:hasKey "THEMA" .

  ?s2 ns2:hasGeometry ?g2.
  ?s2 ns2:hasTag ?tag2.
  ?g2 ns2:asWKT ?wkt2.
  ?tag2 ns2:hasKey "THEMA'" .

  FILTER(FUNCTION(?wkt1, ?wkt2))}
```

Outline

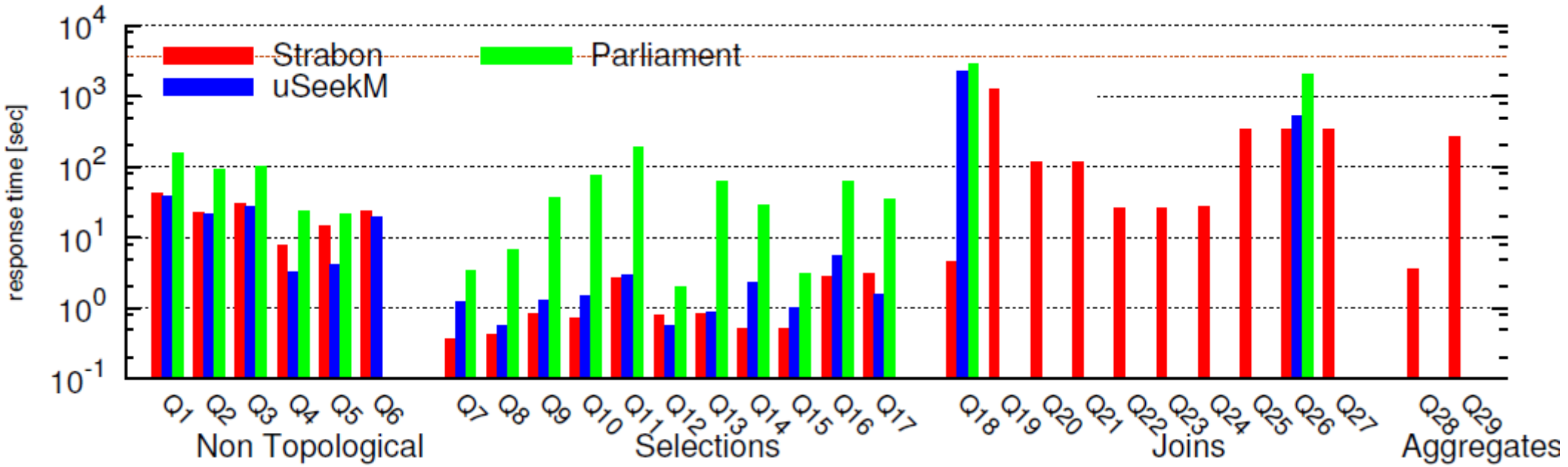
- Motivation
- The benchmark Geographica
  - Real-world workload
  - Synthetic workload
- **Evaluating the performance of geospatial RDF stores using Geographica**
- Conclusions

Experimental Setup

- **Geospatial RDF stores tested:** Strabon, Parliament, uSeekM
- **Machine:** Intel Xeon E5620, 12MB L3 cache, 2.4GHz, 24GB RAM, 4 HDD with RAID-5

- **Micro part (real-world workload) & synthetic workload:**
  - **Metric:** response time
  - Run 3 times and compute the median
  - **Time out:** 1 hour
  - Run both on **warm** caches and **cold** caches
- **Macro part (real-world workload) :**
  - Run **many instantiations** of each scenario for one hour **without cleaning caches**
  - **Metric:** Average time for a complete execution

Results
Real Workload - micro part (cold caches)



Results
Macro part

Scenario	Strabon	uSeekM	Parliament
Reverse Geocoding	65 sec	0.77 sec	2.6 sec
Map Search and Browsing	0.9 sec	0.6 sec	22.2 sec
Rapid Mapping for Fire Monitoring	207.4 sec	-	-

Results

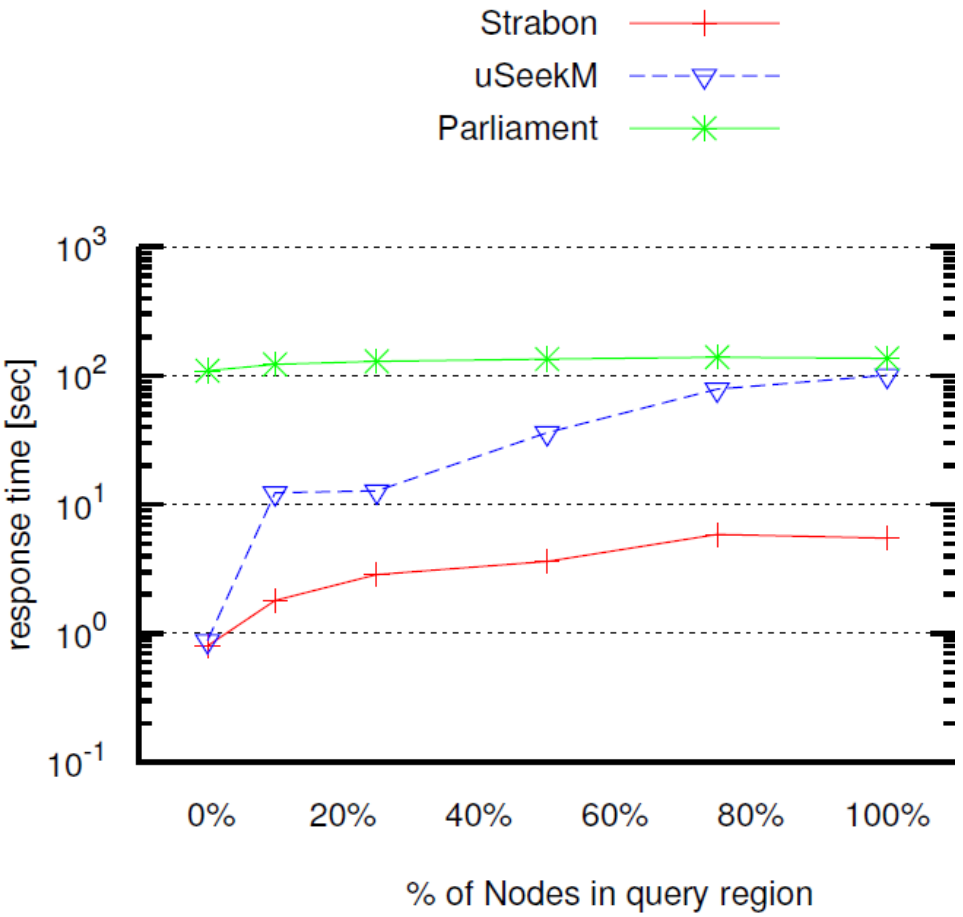
Synthetic Workload

- We generate the synthetic dataset with **n=512**. This results in:
  - 28,900 states
  - 262,144 land ownerships
  - 512 roads
  - 262,144 points of interest

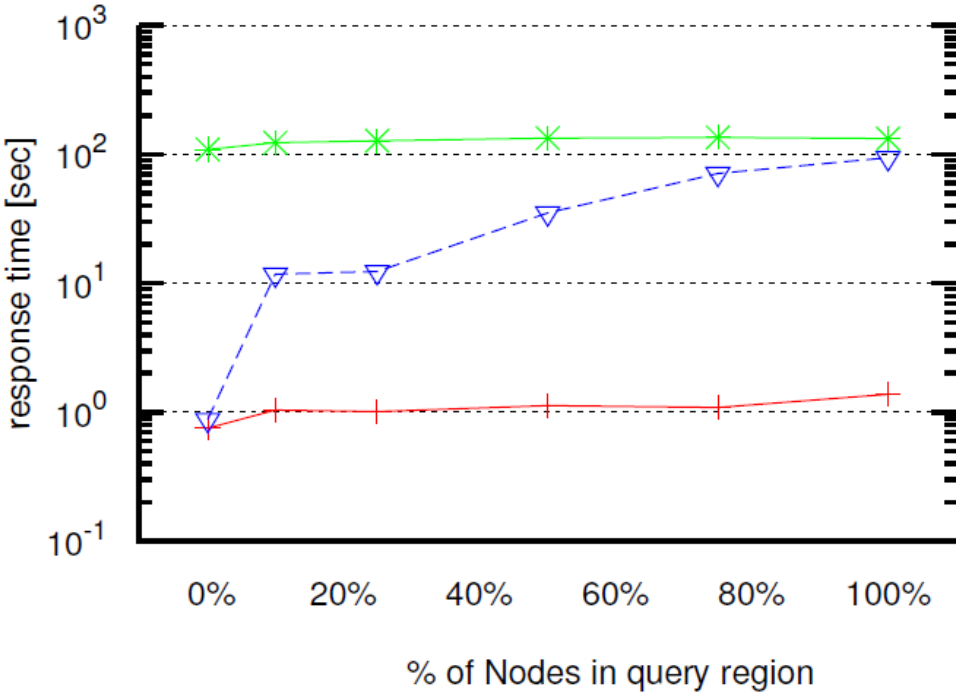
- **Size:** 3,880,224 triples (745 MB)

Results

Synthetic Workload – spatial selections



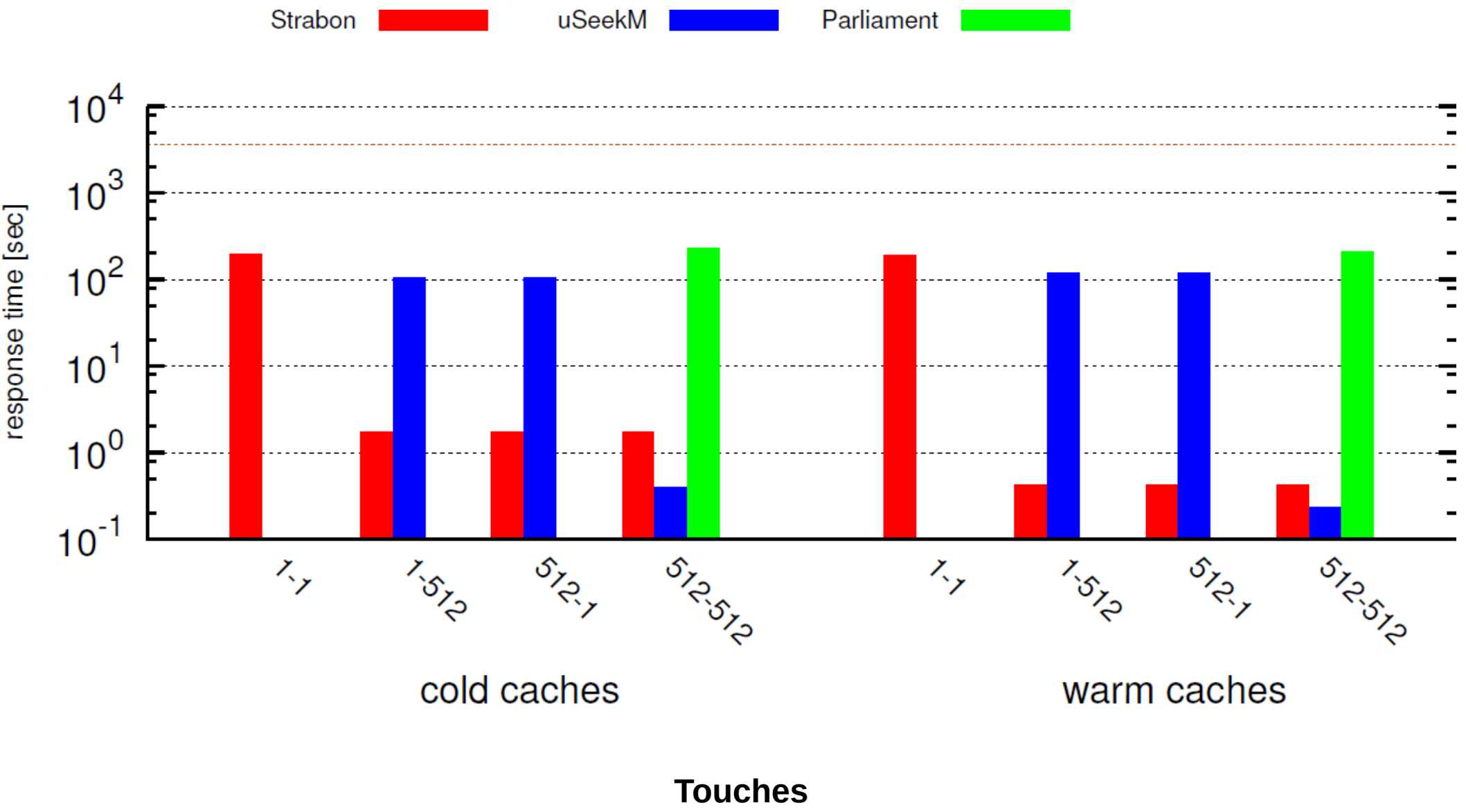
Intersects
Tag 1, cold caches



Intersects
Tag 512, cold caches

Results

Synthetic Workload - Spatial Joins



Conclusions

- We defined **Geographica**, a new comprehensive benchmark for geospatial RDF stores
- Two workloads: **real-world** and **synthetic**
- We used Geographica to compare 3 relevant systems
  - Strabon
  - Parliament
  - uSeekM

Future Work

- Capture the full GeoSPARQL standard.
- Study scaling issues with larger datasets.
- Add more application scenarios
- Extent the generator to produce datasets that do not follow a uniform distribution.
- Extent the benchmark to include time-evolving geospatial data.

Thanks!

- Geographica: **http://geographica.di.uoa.gr**

✉ggarbis@di.uoa.gr

✉Kostis.Kyzirakos@cwi.nl

✉koubarak@di.uoa.gr

- This work was supported in part by the European Commission project **TELEIOS** http://www.earthobservatory.eu

More at ISWC

- **Demo:** SexTant: Visualizing Time Evolving Linked Geospatial Data

Any Questions?