Querying Incomplete Geospatial Information in RDF

Charalampos Nikolaou and Manolis Koubarakis

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens

International Symposium on Spatial and Temporal Databases (SSTD) 2013
August 23, 2013

Motivation

- Increased interest in publishing geospatial datasets as **linked data** (i.e., encoded in RDF and with semantic links to other datasets)

- Geospatial information might be:
  - Quantitative (e.g., exact geometric information)
  - Qualitative (e.g., topological relations)

  ... and express **knowledge** that is

  - Complete
  - Incomplete (or indefinite)

Ordnance Survey (UK)

SO16 0AS

SO16 0AS is a Postcode Unit.

Objects related to "SO16 0AS"

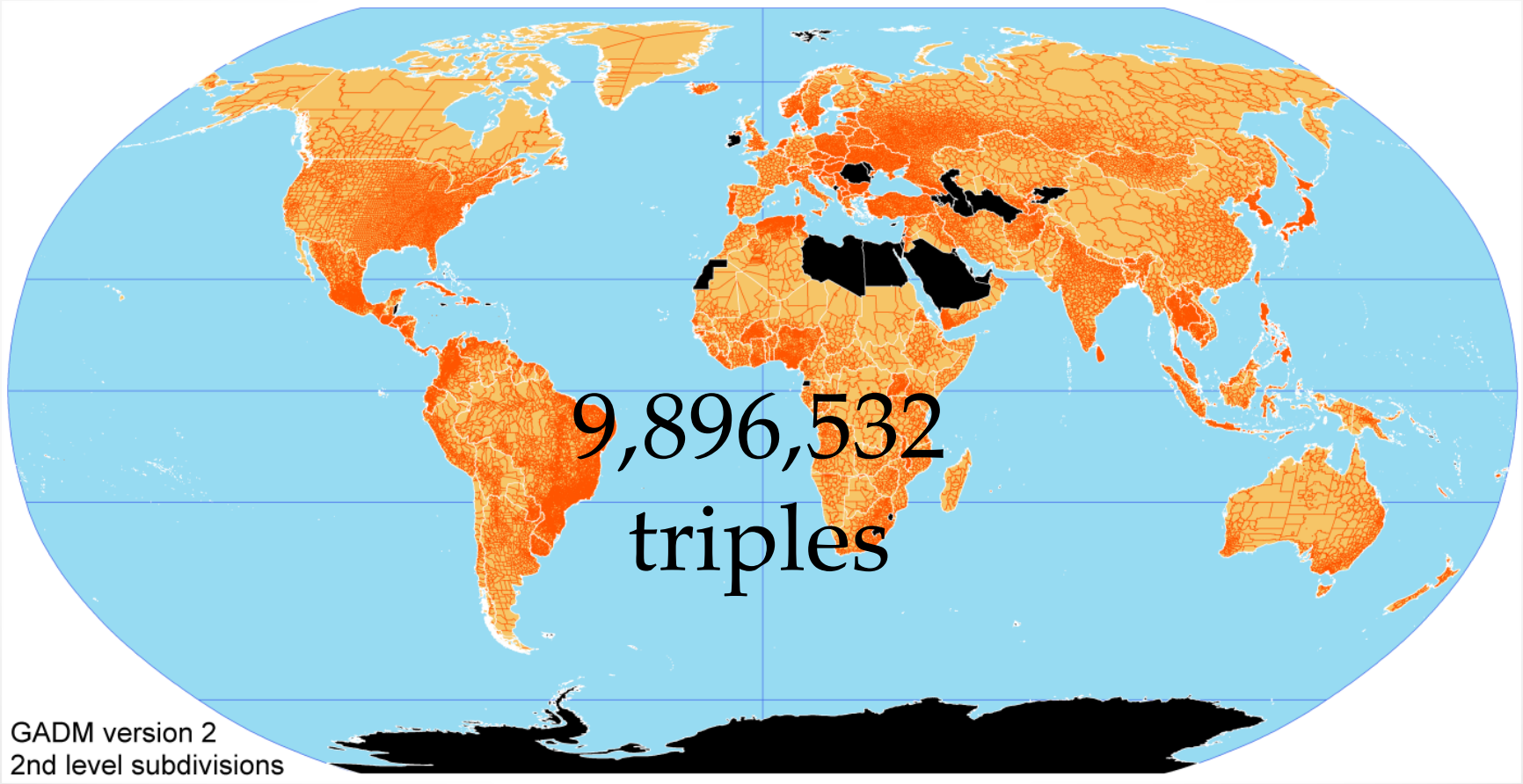
County	Hampshire
Within	SO16 0 SO SO16
NhsHA	E18000009
Nhs RegionalHA	E19000002
Positional Quality Indicator	10

73,546,231 triples

Core facts about "SO16 0AS"

Type	Postcode Unit
Label	SO16 0AS
Northing	115515.00
Easting	437293.00
Lat	50.937872
Long	-1.470616
LH	E18000009
Notation	SO16 0AS

Global Administrative Areas (GADM)



Nomenclature of Territorial Units for Statistics (NUTS)

Home

LV00 - Latvija

This file is a human-readable representation of both the feature and its geometry.

View the feature as: Turtle, RDF/XML

View the geometry as: Turtle, RDF/XML, GML, KML

Feature

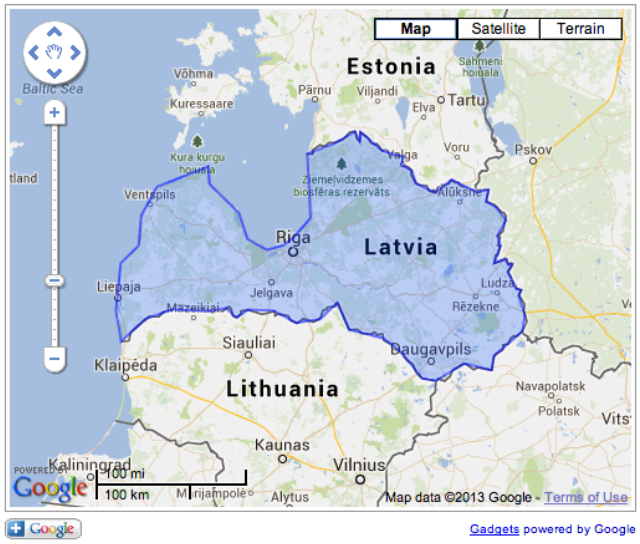
rdf:type	ramon:NUTSRegion, spatial:Feature
rdfs:label	LV00 - Latvija
ramon:name	Latvija
ramon:level	2
ramon:code	LV00
ngeo:geometry	nuts:LV00_geometry
spatial:PP	LV0
spatial:PPi	http://dbpedia.org/resource/Daugavpils
spatial:EC	http://nuts.geovocab.org/id/EE
spatial:EC	http://nuts.geovocab.org/id/LT
spatial:EC	http://nuts.geovocab.org/id/EE0
spatial:EC	http://nuts.geovocab.org/id/LT0
spatial:EC	http://nuts.geovocab.org/id/EE00
spatial:EC	http://nuts.geovocab.org/id/LT00
spatial:EC	http://nuts.geovocab.org/id/EE004
spatial:EC	http://nuts.geovocab.org/id/LT005
spatial:EC	http://nuts.geovocab.org/id/LT006
spatial:EC	http://nuts.geovocab.org/id/LT003
spatial:EC	http://nuts.geovocab.org/id/LT008
spatial:EC	http://nuts.geovocab.org/id/LT009
spatial:EC	http://nuts.geovocab.org/id/EE008
owl:sameAs	http://estatwrap.ontologycentral.com/dic/geo#LV00
owl:sameAs	http://rdfdata.eionet.europa.eu/ramon/nuts2008/LV00
owl:sameAs	http://ec.europa.eu/eurostat/ramon/rdfdata/nuts2008/LV00
owl:sameAs	http://nuts.psi.enakting.org/id/LV00

Geometry

rdf:type	ngeo:Polygon

316,246 triples

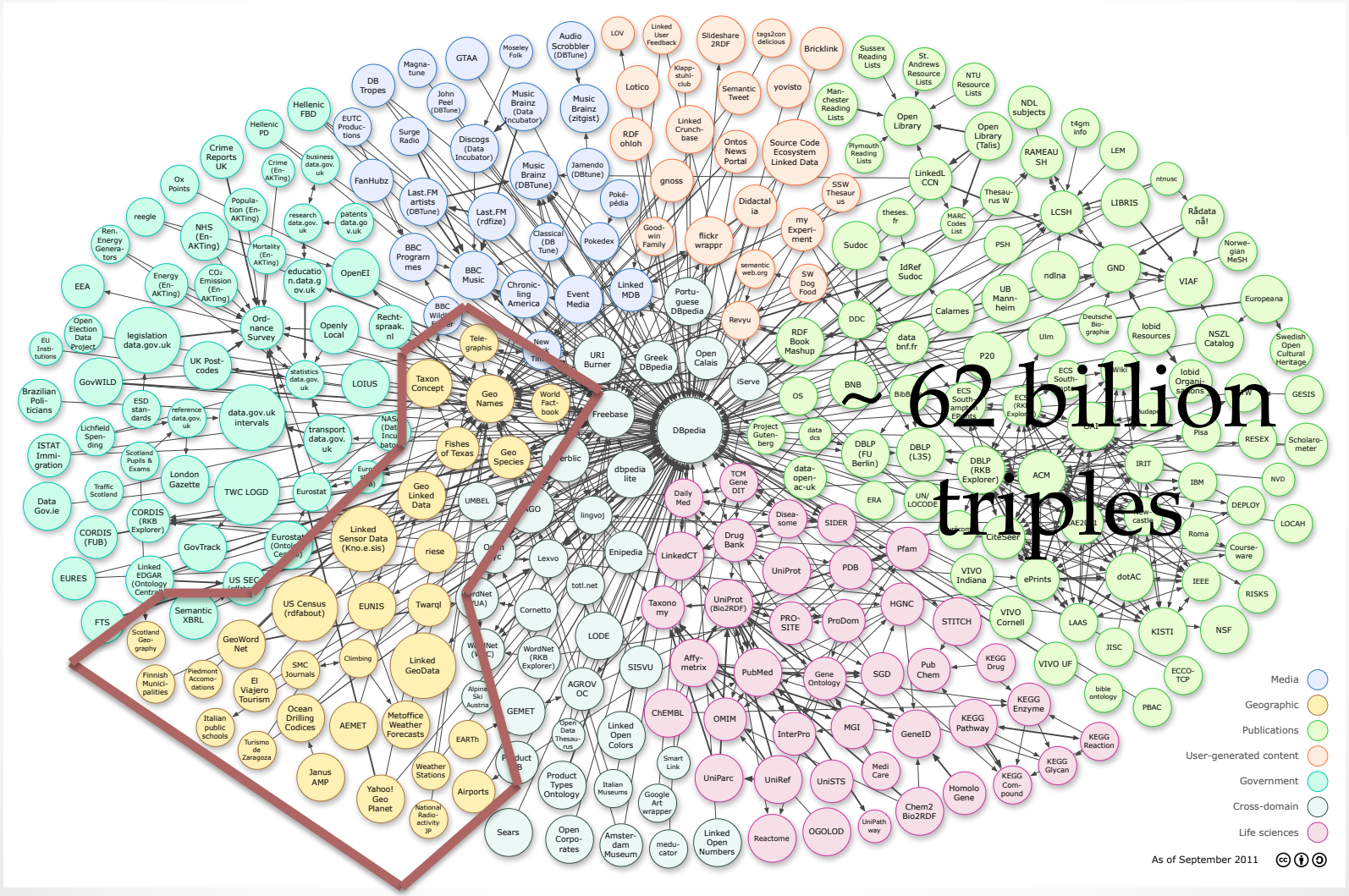
Embedded KML Viewer



Google

Gadgets powered by Google

Linked Geospatial Data



Question

How do we **manage** (represent, store, query) this data efficiently?

Challenges: Theory

① RDF extensions for **representing** and **querying** incomplete qualitative and quantitative geospatial information

- **GeoSPARQL**
  - Standard OGC query language for RDF data with geospatial information
  - Topological relations can be **expressed/queried**, but no reasoning is offered.

- We proposed **RDF^i**
  - Can work with any topological/temporal constraint language with/ without constant symbols (e.g., RCC-5, RCC-8, IA)
  - Formal semantics and algorithm for computing certain answers
  - Preliminary complexity results for various constraint languages

Open issue

- **No** published algorithm for query processing when considering RCC-8 **and** constants

RDF^{i} by example


```
gag:Region          rdfs:subClassOf geo:Feature.
gag:WestGreece      rdf:type gag:Region.
gag:Municipality    rdfs:subClassOf geo:Feature.

gag:OlympiaMuni     rdf:type gag:Municipality.

noa:Hotspot         rdfs:subClassOf geo:Feature.
noa:hotspot         rdf:type noa:Hospot.

noa:Fire            rdfs:subClassOf geo:Feature.
noa:fire            rdf:type noa:Fire.
```

```
gag:OlympiaMuni geo:hasGeometry ex:oGeo.
ex:oGeo rdf:type sf:Polygon.
ex:oGeo geo:asWKT "POLYGON((..))"^^geo:wktLiteral.

noa:hotspot geo:hasGeometry ex:rec.
ex:rec geo:asWKT "POLYGON((..))"^^geo:wktLiteral.

gag:WestGreece geo:sfContains gag:OlympiaMuni.
   noa:hotspot geo:sfContains noa:fire.
```

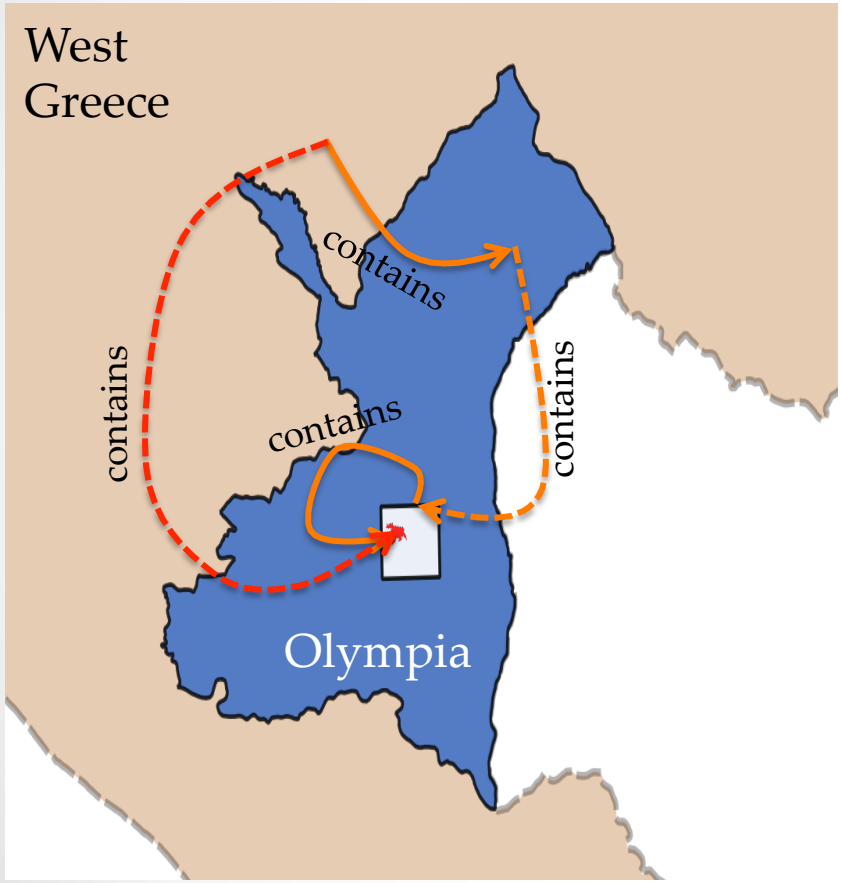
RDF^{i} by example (cont'd)

Query: Find fires inside the region of West Greece.

GeoSPARQL query:

```
CERTAIN SELECT ?f
WHERE {
   ?f rdf:type noa:Fire.
   gag:WestGreece geo:sfContains ?f.
}
```

RDF^{i} by example (cont'd)



Query: Find fires inside the region of West Greece.

GeoSPARQL query:

```
CERTAIN SELECT ?f
WHERE {
   ?f rdf:type noa:Fire.
   gag:WestGreece geo:sfContains ?f.
}
```

Challenges: Theory

② Efficient computation of the entailment relation

$$\Phi \models \Theta$$

- where Φ and Θ are quantifier-free first-order formulas of a constraint language expressing the **topological relations** of various frameworks (RCC-8, DE-9IM, etc.)

Challenges: Theory

③ Computing entailment **is equivalent to** checking consistency of formulas with constraint networks

- Constraint networks:
  - Spatial relations among regions
  - Regions might be **constant** ones (exact geometric information) or identified by a **URI**

- Most recent results considered **basic** and **complete** **RCC-5** networks with **polygonal** regions

Open issue

- For **RCC-8**, deciding consistency is NP-complete
- No published **algorithm** for checking consistency
- Are there **tractable cases**?

Challenges: Practice

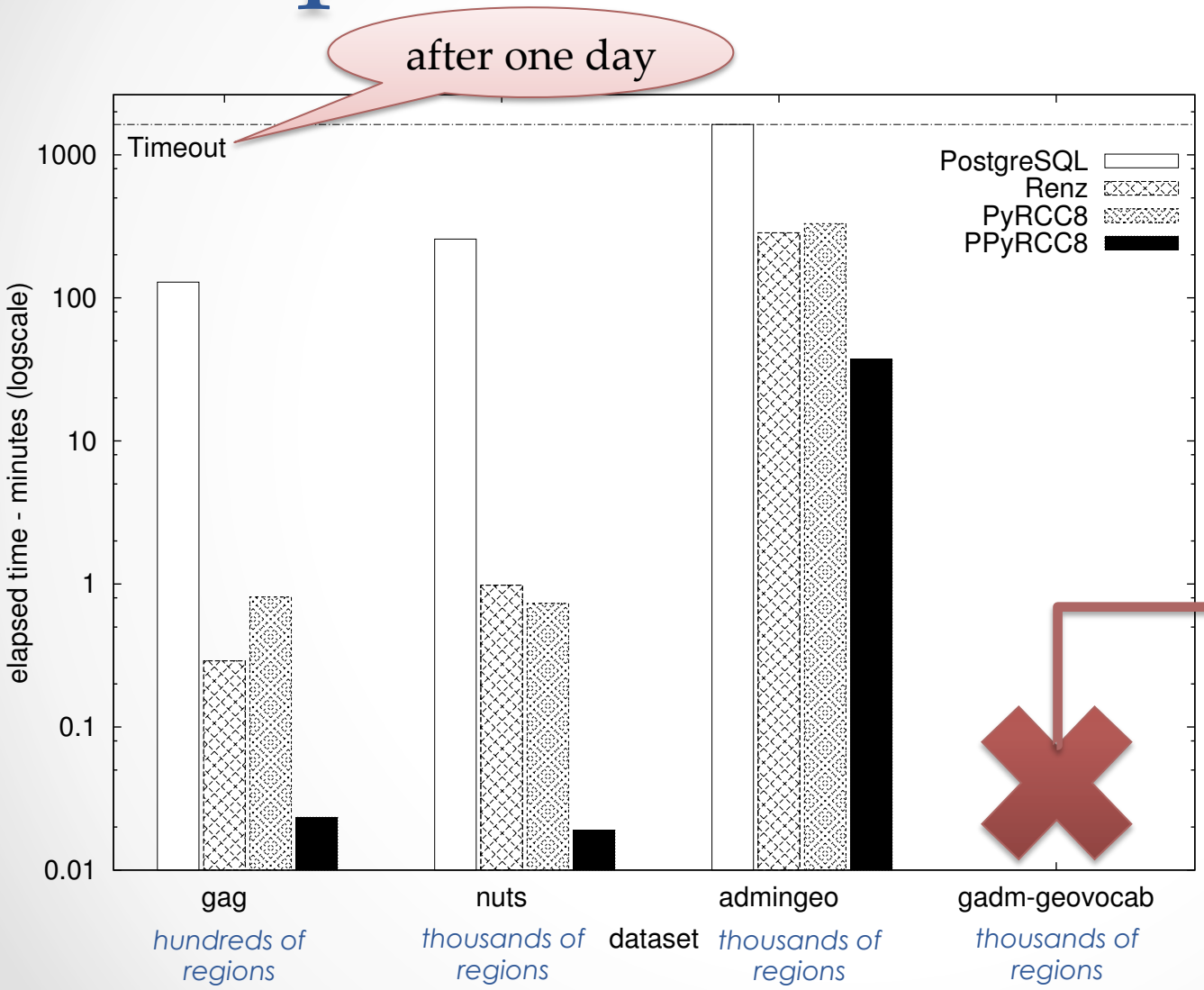
④ Scale to billions of triples

- Reasoners from QSR scale **only up to** hundreds of regions with complex spatial relations

How do they perform in our case?

- Setting:
  - Real linked geospatial datasets
  - No constants
  - Only base RCC-8 relations
  - Evaluation of **consistency checking** using the well-known path-consistency algorithm

Experimental evaluation



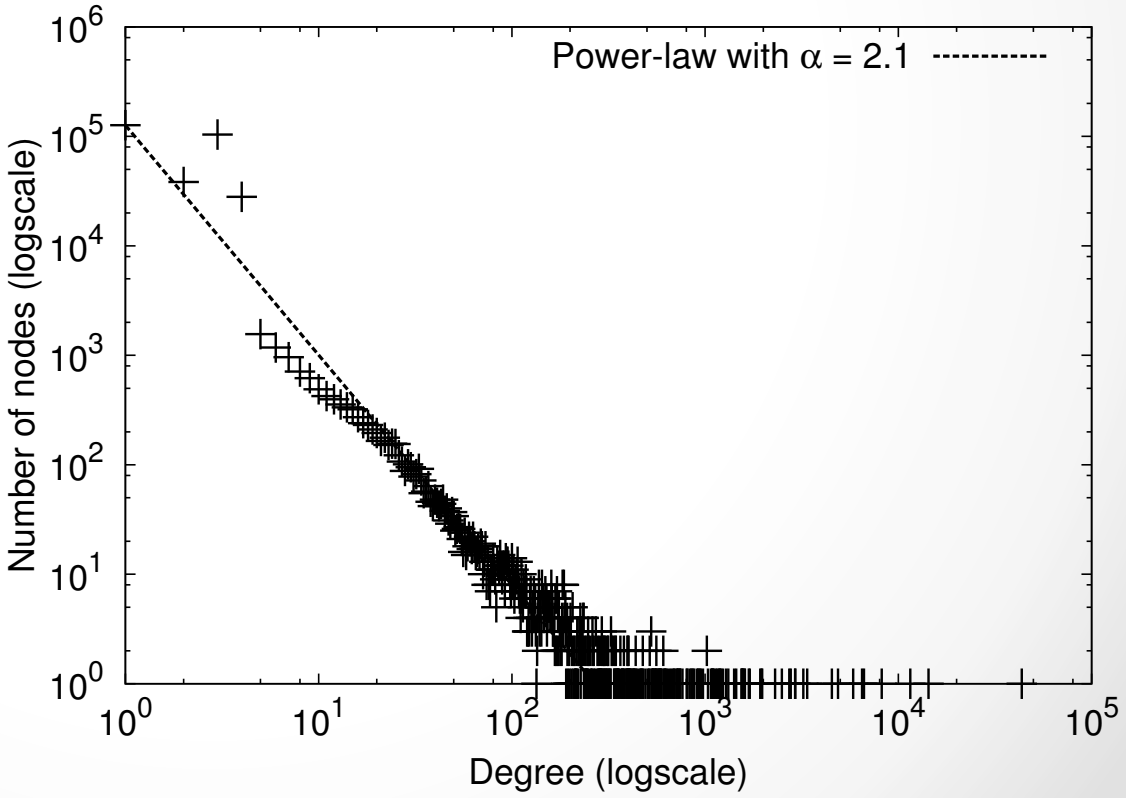
- Computation of the **complete** constraint network
- Running time: $O(n^3)$
- Memory requirements: **$O(n^2)$**

n ≈ thousands to millions

Setup: Intel Xeon E5620, 2.4 GHz, 12MB L3, 48GB RAM, RAID 5, Ubuntu 12.04

Network structure

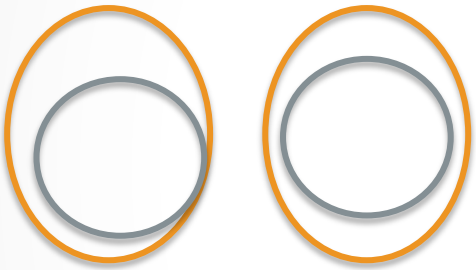
- We have started working on algorithms taking into account the structure of these networks:
  - **Node degrees** fit a power-law distribution
  - Network is sparse



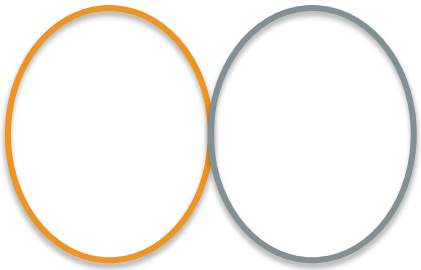
Network structure (cont'd)

- Edges of three kinds:

non-tangential proper part



externally connected



equals

- Reflect networks composed of components with hierarchical structure
  - R-tree extensions (Papadias, Kalnis, Mamoulis, AAAI'99)
- Parallel algorithms combined with backward-chaining techniques for lazy query processing
  - Graph partitioning
  - Path compression data structures and indexes

Related work: Spatial

- Qualitative spatial reasoning
  - Efficient algorithms for **consistency checking** of constraint networks (**complex** spatial relations, **few** number of regions)
  - Does not consider **query processing**

- Description logic reasoners
  - **PelletSpatial**: RCC-8 reasoning (cannot handle disjunctions)
  - **RacerPro**: RCC-8 reasoning

Related work: Temporal

- Chaudhuri (VLDB'88)
- The knowledge representation language Telos (TOIS'90)
- Foundations of temporal constraint databases (Koubarakis, PhD thesis, '94)
- Qualitative temporal reasoning community (since 80s)
- SQL+i system (BNCOD'96)
- Later system (IEEE'97)
- Hurtado and Vaisman (2006)

Conclusions

- What's the **CHALLENGE**?

Implementing an **efficient** query processing system for incomplete geospatial information in RDF[i]

- **The desired system** should:
  - **reason** about **qualitative** and **quantitative** spatial information that might be **incomplete**
  - be **scalable** to **billions** of triples in the most useful cases

Thank you

Dataset characteristics

Dataset	#triples	#regions	#RCC-8 relations
ADMGB	149 046	11 762	77 907
GAG	11 780	412	3023
NUTS-RDF	316 246	2236	3176
GADM-RDF	9 896 532	276 728	590 445
GADM-RDF-EUROPE	355 656	23 037	51 309