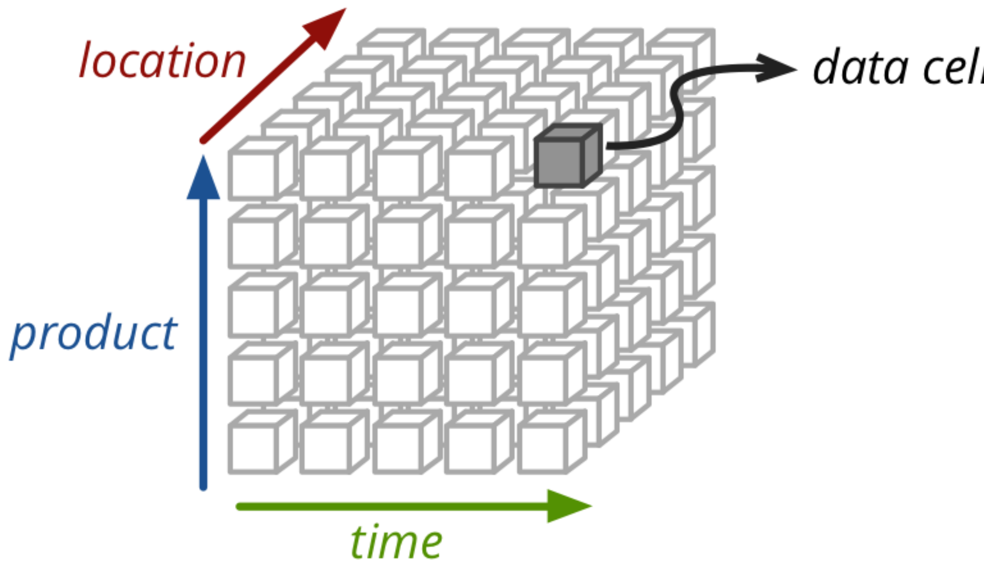
Containment and Complementarity Relationships in Multidimensional Linked Open Data

Marios Meimaris and George Papastefanatos
Institute for the Management of Information Systems
Research Center "Athena"

{m.meimaris, gpapas}@imis.athena-innovation.gr

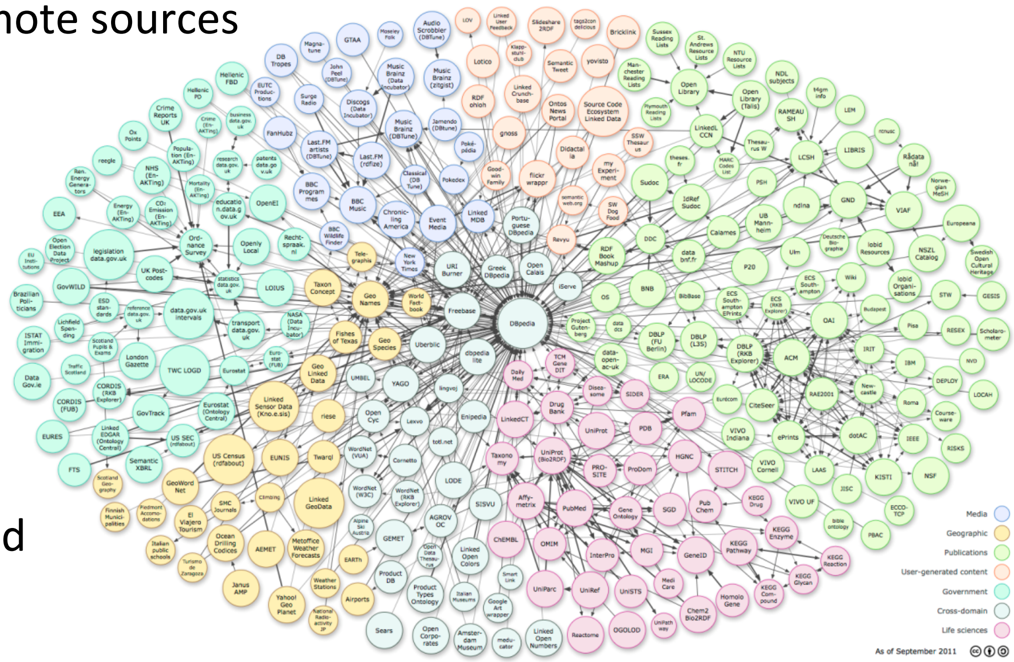
Multidimensional data

- Schema
  - Dimensions
  - Measures
  - Attributes
  - Code lists
- Data
  - Observations



Multidimensional Linked Data

- Origin of different source datasets
- LD recommendations and Best Practices provide **common grounds** across remote sources
- RDF Data cube[1] provides a common meta-schema
- Re-use of:
    - Dimension properties
    - Measure properties
    - Code lists
    - *Hierarchies*
- In case of no re-use, mapping/alignment is needed



1. http://www.w3.org/TR/vocab-data-cube/

Problem tackled

- Relating points in multidimensional data spaces semantically
- Bulk detection and computation of **containment** and **complementarity** relationships between observations
  - in the same dataset or
  - in different datasets
- Observation relationships are useful for:
  - performing OLAP analytics over multidimensional, multi-dataset data spaces
  - computing similarities/distances between observations
  - Suggestion mechanisms for relevant statistics
  - Exploratory analysis and discovery

Observations are related

- We identify two (non-exhaustive) types of relationships:
  - Observation *containment*
  - Observation *complementarity*

D_1

	refArea	refPeriod	sex	ex:Population
o_{11}	Athens	2001	Total	5M
o_{12}	Austin	2011	Male	5.5M

D_2

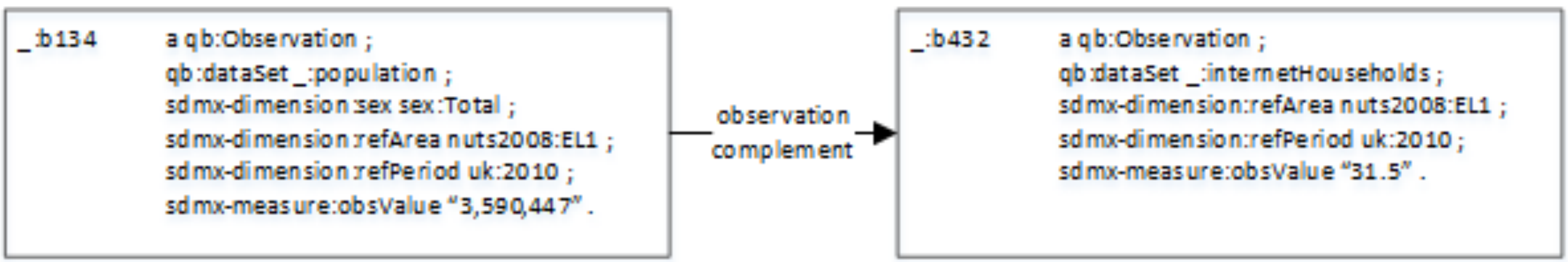
	refArea	refPeriod	ex:Unemployment	ex:Poverty
o_{21}	Greece	2011	26%	15%
o_{22}	Italy	2011	20%	10%

D_3

	refArea	refPeriod	ex:Unemployment
o_{31}	Athens	2001	10%
o_{32}	Athens	Jan2011	30%
o_{33}	Rome	Feb2011	7%

Observation Complementarity

- Two observations complement each other when they provide different information for the same point in the data space



$$(P_a \subseteq P_b) \wedge \left(\forall p_i \in P_a \cap P_b : h_a^i = h_b^i\right) \wedge \left(\forall p_j \in P_b \backslash P_a : h_b^j = c_{jroot}\right)$$

P_k: the set of dimension properties for observation l

p_i: a single dimension property

h_l^m : the value of property m for observation l

c_{jroot}: the top (root) concept for all hierarchies

Observation Containment

- An observation contains another observation when it is a *partial* or *full* generalization of the latter w.r.t to their shared dimension values
- *Full* containment vs *Partial* containment
  - Full containment means that a contained/containing observation can be directly rolled-up/drilled-down to the containing/contained observation,
  - Partial containment means that both contained and containing observation must be *rolled-up on their disjoint dimensions* to complement each other

full $(\exists M_i \in M_a \cap M_b) \wedge (P_a \subseteq P_b) \wedge (\forall p_i \in P_a \cap P_b : h_a^i \vdash h_b^i)$

partial $(\exists M_i \in M_a \cap M_b) \wedge (P_a \subseteq P_b) \wedge (\exists p_i \in P_a \cap P_b : h_a^i \vdash h_b^i)$

Containment example

	location	time	sex	Population
obs1	Italy	2012	Total	59,478,000
obs2	Riva del Garda	2012	Male	15,100
obs3	Trentino	2012	Female	248,400
	…	…	…	…

full
partial

Hierarchy is reflexive (i.e. a value is a parent of itself)

Computation

1. Build the feature space
2. Group by dimension / measure
3. Extract containment per dimension / measure
4. Compute overall containment scores and classify as full or partial
5. Compute complementarity scores

Occurrence Matrix

1. Build the feature space into an **occurrence matrix**
   - Each dimension value is a feature
   - Encoded is the hierarchy of features (1 for occurrence and all parents, 0 otherwise)

	refArea										refPeriod					sex		
	WLD	EUR	AM	GR	IT	Ath	Rom	US	TX	Aus	ALL	2001	2011	Jan11	Feb11	M	F	T
obs_{11}	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	1
obs_{12}	1	0	1	0	0	0	0	1	1	1	1	0	1	0	0	1	0	1
obs_{21}	1	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1
obs_{22}	1	1	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	1
obs_{31}	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	1
obs_{32}	1	1	0	1	0	1	0	0	0	0	1	0	1	1	0	0	0	1
obs_{33}	1	1	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	1

ΙΠΣΥ IMIS

Athena Research Center
Research and Innovation Center in Information, Communication and Knowledge Technologies

Containment Matrices

2. For N observations, compute one NxN containment matrix $\mathbf{CM}_{\mathbf{pm}}$ for each dimension p_m in the set of all datasets. Then cell [i,j] becomes:

- 1 if values of dimension are parent-child for observations i and j, or
- 0 otherwise

Function *sf* to determine this for observations o_a and o_b and dimension p_m:

$$sf(o_a, o_b)|_{p_m} = \begin{cases} 1, & (a\ AND\ b) = b \\ 0, & otherwise \end{cases}$$

where a and b are the bit vectors of observations

Containment relationships

3. Adding all containment matrices $\mathbf{CM_{pm}}$ yields *full* and *partial* containment relationships in an overall containment matrix **OCM**:

$$\boldsymbol{OCM} = \sum_{i=1}^{k} u_i\, CM_i \Big/ \sum_{i=1}^{k} u_i$$

For observations o_a and o_b:

- o_a cont_{full} o_b *iff* $\mathbf{OCM}[o_a, o_b]=1$
- o_a cont_{part} o_b *iff* $0 < \mathbf{OCM}[o_a, o_b] < 1$

Complementarity relationships

4. Complementarity is computed as follows:

$$cf(o_a, o_b) = \begin{cases} 1, & (sf(o_a, o_b)|_P = 1) \text{ AND } (a = b) \\ 0, & otherwise \end{cases}$$

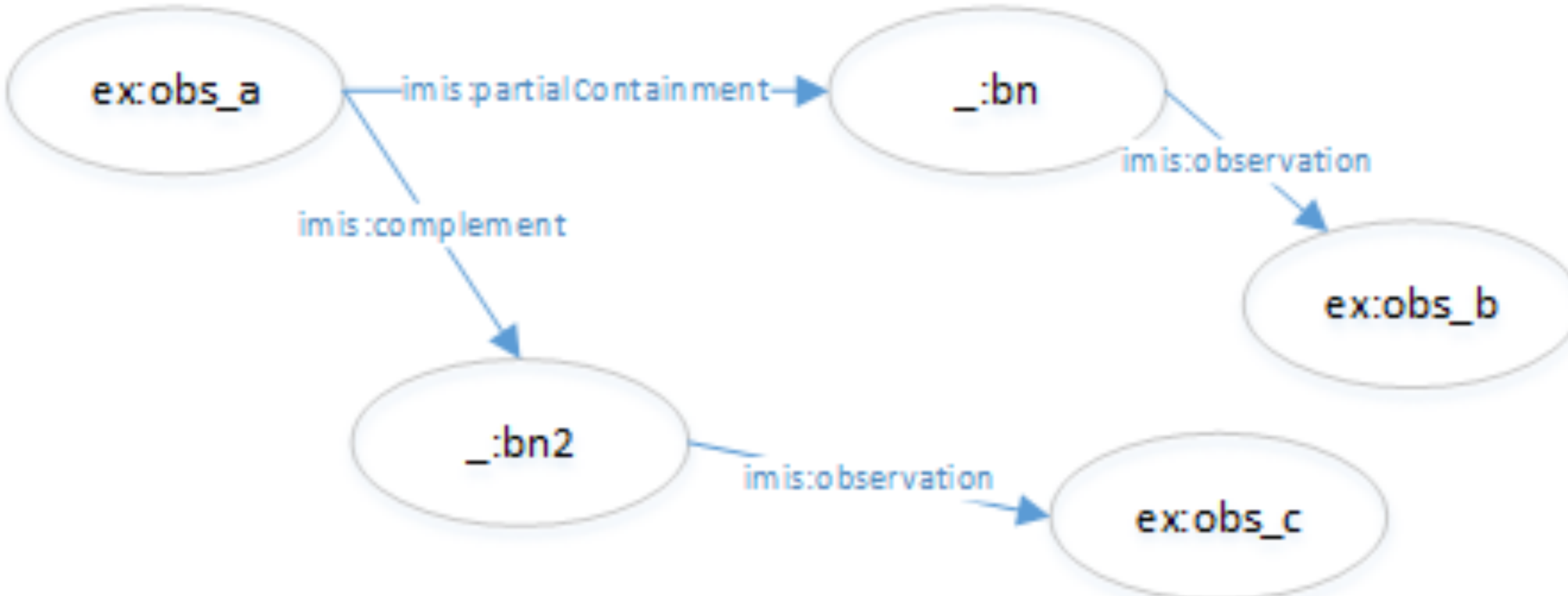
where P the occurrences of dimension properties and a, b the bit vectors of o_a and o_b in the occurrence matrix

For observations o_a and o_b:

- o_a *compl*$_{full}$ o_b *iff* **OCM**$[o_a, o_b] > 0$

Containment is transitive, complementarity is symmetric

Data Cube Extension



Athena Research Center
Research and Innovation Center in Information, Communication and Knowledge Technologies

Experimental Evaluation

- Datasets:
    - Population (Eurostat, Worldbank)
    - Internet households (Eurostat)
    - Poverty (Eurostat, Worldbank)
- 6 dimension properties
- 3 measure properties

#of obs.	refArea	refPeriod	sex	unit	age	poverty	internet	population
D_1 (539)	85 regions,20 countries	2004-2011	N/A	Yes	Yes	Yes	N/A	N/A
D_2 (1693)	293 regions, 33 countries	2003-2010	N/A	Yes	N/A	Yes	N/A	N/A
D_3 (629)	42 regions, 3 countries	2009-2013	M, F, Total	Yes	N/A	N/A	N/A	Yes
D_4 (316)	65 regions,7 countries	2009-2013	N/A	N/A	N/A	N/A	Yes	N/A

Results - Discussion

- Most new relationships are partial containments (~27% of possible relationships)
- Complementarity is the strictest relationship (0.03% of the total possible observation pairs)
- Relatedness of complementarity to partial/full containment
- ~1.3 million new links between observations

	D_1	D_2	D_3	D_4
D_1	647 (0.31%) full 34.3k (16.32%) partial N/A compl	N/A full N/A partial N/A compl	N/A full N/A partial N/A compl	N/A full N/A partial N/A compl
D_2	605 (0.02%) full 605k (14.83%) partial 1238 (0.04%) compl	3370 (0.14%) full 378k (14.83%) partial N/A (complement	N/A full N/A partial 204 (0.004%) compl	N/A full N/A partial N/A compl
D_3	N/A full N/A partial N/A compl	N/A full N/A partial N/A compl	1k (0.26%) full 261k (65.9%) partial N/A compl	N/A full N/A partial N/A compl
D_4	N/A full N/A partial 328 (0.05%) compl	N/A full N/A partial 218 (0.005%) compl	N/A full N/A partial 592 (0.07%) compl	437 (0.17%) full 22.2k (22.3%) partial N/A compl

Athena Research Center
Research and Innovation Center in Information, Communication and Knowledge Technologies

Future Work

- Suggestion mechanisms based on computed relationships, conduct user studies to evaluate
- Faster and more efficient computations (now $O(N^2)$)
  - Better feature extraction
  - Dimensionality reduction
- Extracting *latent datasets* based on containment and complementarity relationships

Support

- **DIACHRON**

  Managing the Evolution and Preservation of the Data Web

- **KRIPIS: SODAMAP Project**

regions at the centre of development

Co - financed by the Hellenic Republic and the European Union - European Regional Development Fund,
in the context of the O.P. Competitiveness and Entrepreneurship (OPC II) and the R.O.P. Attica, R.O.P. Macedonia - Thrace

- **linked-statistics.gr**