



**HAL**  
open science

## Structural Summarization of Semantic Graphs

Ioana Manolescu

► **To cite this version:**

Ioana Manolescu. Structural Summarization of Semantic Graphs. Extended Semantic Web Conference (ESWC) , Jun 2018, Heraklion, Greece. . hal-01808737

**HAL Id: hal-01808737**

**<https://inria.hal.science/hal-01808737v1>**

Submitted on 6 Jun 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Structural Summarization of Semantic Graphs

Ioana Manolescu

Head of CEDAR team

INRIA and Ecole Polytechnique, France

`ioana.manolescu@inria.fr`

`http://pages.saclay.inria.fr/Ioana.Manolescu`

ESWC Conference, June 5, 2018

# Outline

- 1 **Motivation:** **data discovery** in semantic-rich RDF graphs
- 2 **Framework:** **quotient summaries**
  - Smaller graph which represents the original one in some sense
- 3 **Proposal:** use **property cliques** for summarizing **explicit** and **implicit** data
- 4 Summarization **algorithms**
- 5 Perspectives

Joint work with François Goasdoué (U. Rennes 1 and Inria), Paweł Guzewicz (U. Paris Saclay and Inria), and Šejla Čebirić (U. Paris Saclay and Inria) [ČGM15a, ČGM15b, ČGM17a, GM18, PGA<sup>+</sup>18]

# Part I

Motivation: data discovery in RDF graphs

# Big Data needs semantics

AI Magazine, Spring 2015



The image shows two side-by-side screenshots of the Data.gov website's search results page. Both screenshots are for the 'Data Catalog' section, with the top navigation bar including 'DATA TOPICS - IMPACT APPLICATIONS DEVELOPERS CONTACT'.

The left screenshot shows search results for 'Natural Disaster'. The search bar contains 'Natural Disaster' and the results are ordered by 'Relevance'. The main heading is '93 datasets found for "Natural Disaster"'. Below this, there are several dataset cards with green arrows pointing to the right, indicating they are highlighted. The cards include:
 

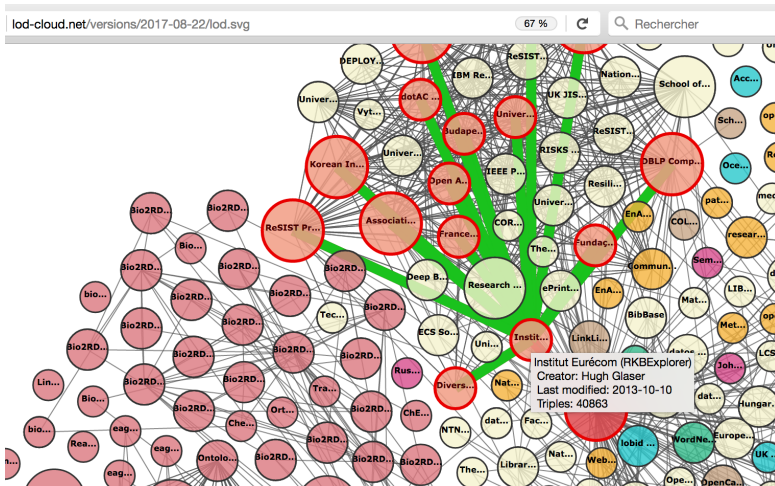
- FEMA Disaster Declarations Summary**: Federal Emergency Management Agency Department of Homeland Security - FEMA Disaster Declarations Summary is a summarized dataset describing all Federally Declared Disasters. This information begins with the first disaster declaration.
- Northeast Crop Disaster Assistance Program**: Department of Agriculture - USDA's Farm Service Agency's (FSA) Northeast Crop Disaster Assistance Program (NAP) provides financial assistance to producers of nonperennial crops when they...
- Child Nutrition Programs Disaster Response Menu**: Department of Agriculture - This menu option provides an overview of what State agencies, School Food Authorities (SFA) participating in the National School Lunch and School Breakfast...

The right screenshot shows search results for 'Earthquakes'. The search bar contains 'Earthquakes' and the results are ordered by 'Relevance'. The main heading is '243 datasets found for "Earthquakes"'. Below this, there are several dataset cards with green arrows pointing to the right, indicating they are highlighted. The cards include:
 

- Earthquake Feeds**: US Geological Survey Department of the Interior - Near real-time earthquake information for a variety of time windows in a variety of formats.
- Earthquake Locations**: State of North Dakota - This layer has been compiled from real-time sources depicting the locations of earthquakes that have been confirmed to have occurred within the state of North Dakota.
- Earthquake Damage - General**: National Oceanic and Atmospheric Administration, Department of Commerce - An earthquake is the sudden or breaking of the ground (usually by sudden displacement of rock in the Earth's crust). Earthquakes result from crustal stresses...

# RDF graph discovery

An RDF graph can be large and complex, lack a fixed schema, include many heterogeneous values...



# RDF summaries

## **Simplified views** of an RDF graph [ČGK<sup>+</sup>18]

- Most often, a summary is also a **graph**, and/or: **statistics**, **patterns**...
- Summarize: the **data**, the **ontology**, **both**
- Many prior works on graph summarization applied to RDF

# RDF summaries

## **Simplified views** of an RDF graph [ČGK+18]

- Most often, a summary is also a **graph**, and/or: **statistics, patterns...**
- Summarize: the **data**, the **ontology**, **both**
- Many prior works on graph summarization applied to RDF

Summary uses:

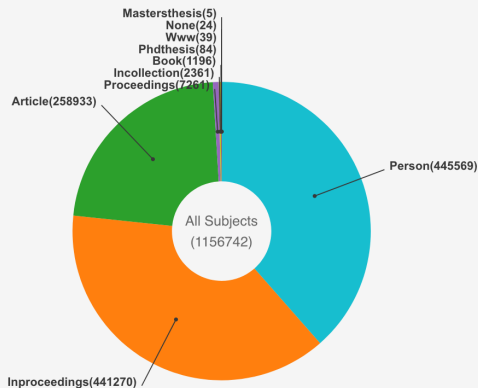
- ① For **query processing**: give direct access to a group of nodes summarized together, detect empty queries...
- ② For **data discovery**: help identify interesting structure or patterns in the data



# RDF graphs are often structurally heterogeneous

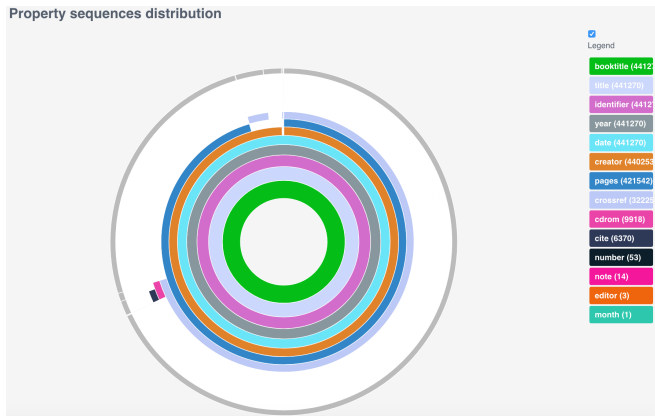
Subject types in DBLP bibliographic data:

**Type distribution** (Click *All Subjects* or a certain type below for further exploration.)



# RDF graphs are often structurally heterogeneous

Data properties of DBLP conference articles:



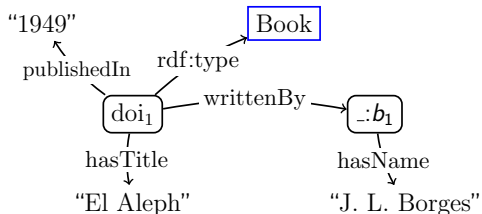
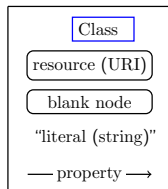
# Our goal

Define **structural summaries** which:

- resist to **heterogeneity**
- flexibly take into account **RDF types**
- allow summarizing **implicit triples**
- can be **implemented efficiently**

# The Resource Description Framework (RDF)

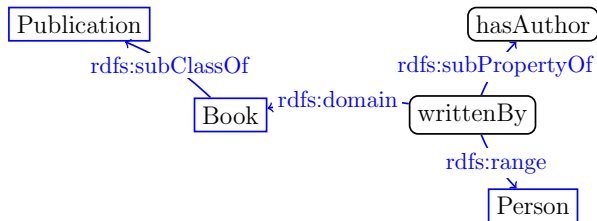
**RDF graph:** set of triples



# RDF Schema

We consider **RDFS** deductive constraints, stating connections between classes and properties

| Constraint      | Triple                         | OWA interpretation                   |
|-----------------|--------------------------------|--------------------------------------|
| <b>Subclass</b> | $c_1$ rdfs:subClassOf $c_2$    | $c_1 \subseteq c_2$                  |
| Subproperty     | $p_1$ rdfs:subPropertyOf $p_2$ | $p_1 \subseteq p_2$                  |
| Domain typing   | $p$ rdfs:domain $c$            | $\Pi_{\text{domain}}(p) \subseteq c$ |
| Range typing    | $p$ rdfs:range $c$             | $\Pi_{\text{range}}(p) \subseteq c$  |

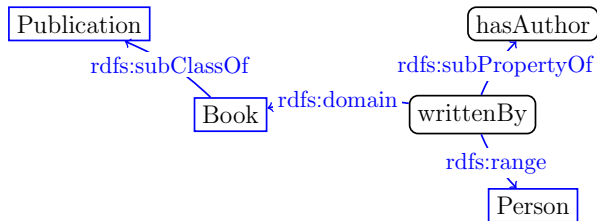


“Any  $c_1$  is also a  $c_2$ ”

# RDF Schema

Simple language of deductive constraints between classes and properties

| Constraint    | Triple                         | OWA interpretation                   |
|---------------|--------------------------------|--------------------------------------|
| Subclass      | $c_1$ rdfs:subClassOf $c_2$    | $c_1 \subseteq c_2$                  |
| Subproperty   | $p_1$ rdfs:subPropertyOf $p_2$ | $p_1 \subseteq p_2$                  |
| Domain typing | $p$ rdfs:domain $c$            | $\Pi_{\text{domain}}(p) \subseteq c$ |
| Range typing  | $p$ rdfs:range $c$             | $\Pi_{\text{range}}(p) \subseteq c$  |

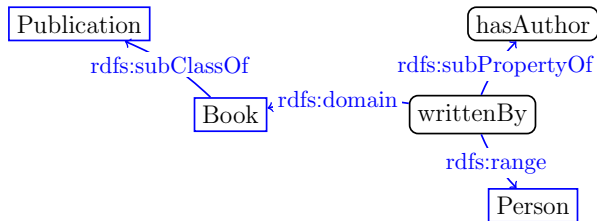


“If two resources are related by  $p_1$ , they are also related by  $p_2$ ”

# RDF Schema

Simple language of deductive constraints between classes and properties

| Constraint    | Triple                         | OWA interpretation                   |
|---------------|--------------------------------|--------------------------------------|
| Subclass      | $c_1$ rdfs:subClassOf $c_2$    | $c_1 \subseteq c_2$                  |
| Subproperty   | $p_1$ rdfs:subPropertyOf $p_2$ | $p_1 \subseteq p_2$                  |
| Domain typing | $p$ rdfs:domain $c$            | $\Pi_{\text{domain}}(p) \subseteq c$ |
| Range typing  | $p$ rdfs:range $c$             | $\Pi_{\text{range}}(p) \subseteq c$  |

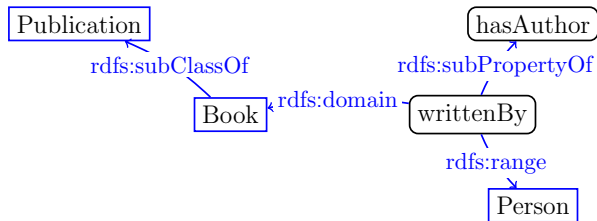


“Anyone having  $p$  is a  $c$ ”

# RDF Schema

Simple language of deductive constraints between classes and properties

| Constraint    | Triple                         | OWA interpretation                   |
|---------------|--------------------------------|--------------------------------------|
| Subclass      | $c_1$ rdfs:subClassOf $c_2$    | $c_1 \subseteq c_2$                  |
| Subproperty   | $p_1$ rdfs:subPropertyOf $p_2$ | $p_1 \subseteq p_2$                  |
| Domain typing | $p$ rdfs:domain $c$            | $\Pi_{\text{domain}}(p) \subseteq c$ |
| Range typing  | $p$ rdfs:range $c$             | $\Pi_{\text{range}}(p) \subseteq c$  |



“Anyone who is a value of  $p$  is a  $c$ ”



# Open-world assumption and RDF entailment

RDF data model based on the open-world assumption.

Deductive constraints lead to **implicit triples**:  
part of the graph even though not explicitly present

# Open-world assumption and RDF entailment

RDF data model based on the open-world assumption.

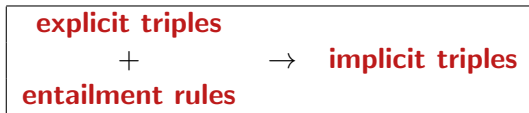
Deductive constraints lead to **implicit triples**:  
part of the graph even though not explicitly present

|                         |   |                         |   |                         |
|-------------------------|---|-------------------------|---|-------------------------|
| <b>explicit triples</b> | + |                         | → | <b>implicit triples</b> |
|                         |   | <b>entailment rules</b> |   |                         |

# Open-world assumption and RDF entailment

RDF data model based on the open-world assumption.

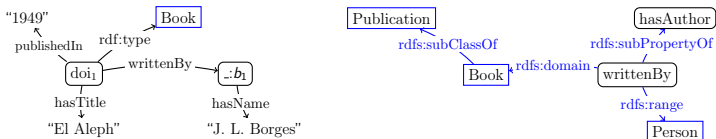
Deductive constraints lead to **implicit triples**:  
part of the graph even though not explicitly present



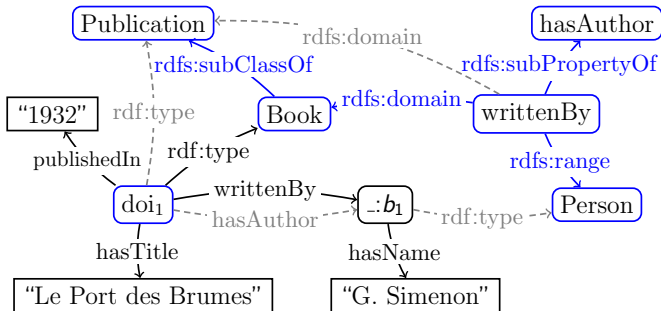
Exhaustive application of entailment leads to **saturation (closure)**

# The semantics of an RDF graph $G$ is its saturation $G^\infty$

RDF data graph and RDF schema graph:



Saturation of the graph union:



## Part III

# RDF summarization

# RDF summaries

## Problem

RDF graph  $G$  is large, heterogeneous, partially implicit.  
How to compactly represent all its structure?

## Existing solutions

**Partial** representation (frequent patterns, statistics etc.)  
e.g., [NM11, LYL13]

**Potentially not compact** e.g., [GW97, CFKP15]  
Only for **explicit data**, e.g., [CDT13, ZDYZ14]

# A summary of DBLP data

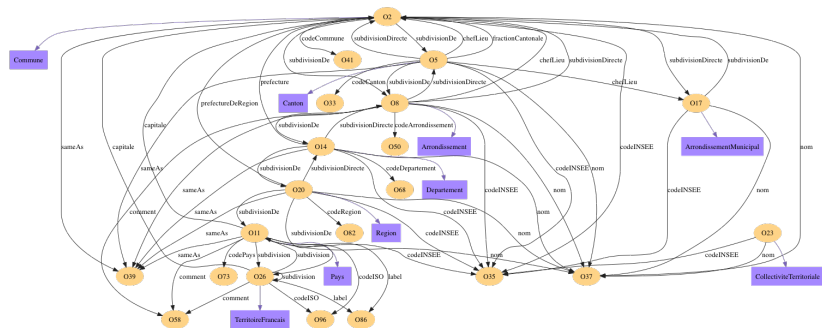
150M triples



# A summary of geographic data

French territory division in regions, departments, urban areas, cities, districts etc.

368K triples



Dataset: <http://inec-gis.fr>. Number of triples: 368417  
 Nodes: 30 (Typed: 9, Untyped: 21, Property: 0)  
 Edges: 69 (Data edges: 69, Schema edges: 0)



## Summarization principle: quotient graphs

Let  $\equiv$  be an equivalence relation on the nodes of  $G$ .

The **quotient  $G_{\equiv}$  of a directed graph  $G$  by  $\equiv$**  is a graph defined as follows:

- $G_{\equiv}$  nodes: one for  $\equiv$  equivalence class of  $V$
- $G_{\equiv}$  edges:  $n_{\equiv}^1 \xrightarrow{a} n_{\equiv}^2$  iff  $\exists n_1 \xrightarrow{a} n_2 \in G$  such that  $n_1$  represented by  $n_{\equiv}^1$ ,  $n_2$  represented by  $n_{\equiv}^2$

## Summarization principle: quotient graphs

Let  $\equiv$  be an equivalence relation on the nodes of  $G$ .

The **quotient  $G_{\equiv}$  of a directed graph  $G$  by  $\equiv$**  is a graph defined as follows:

- $G_{\equiv}$  nodes: one for  $\equiv$  equivalence class of  $V$
- $G_{\equiv}$  edges:  $n_{\equiv}^1 \xrightarrow{a} n_{\equiv}^2$  iff  $\exists n_1 \xrightarrow{a} n_2 \in G$  such that  $n_1$  represented by  $n_{\equiv}^1$ ,  $n_2$  represented by  $n_{\equiv}^2$

Quotients have interesting summary qualities:

- 1 **Property completeness:** All  $G$  properties appear in  $G_{\equiv}$
- 2 **Size guarantees:** By definition,  $G_{\equiv}$  is at most as large as  $G$  (usually much smaller)
- 3 **Structure representativeness:** Given a query  $q$ , if its **structure-only** version is empty on  $G_{\equiv}$ , then  $q$  is empty on  $G$

## Common graph quotients: bisimilarity [HHK95]

Two nodes are forward (resp. backward) bisimilar if they have exactly the same incoming (resp. outgoing) paths;  $\sim_{fw}$ ,  $\sim_{bw}$ ,  $\sim_{fb}$

## Common graph quotients: bisimilarity [HHK95]

Two nodes are forward (resp. backward) bisimilar if they have exactly the same incoming (resp. outgoing) paths;  $\sim_{fw}$ ,  $\sim_{bw}$ ,  $\sim_{fb}$

**Problem:** Bisimilarity compresses/summarizes very little!

## Common graph quotients: bisimilarity [HHK95]

Two nodes are forward (resp. backward) bisimilar if they have exactly the same incoming (resp. outgoing) paths;  $\sim_{fw}$ ,  $\sim_{bw}$ ,  $\sim_{fb}$

**Problem:** Bisimilarity compresses/summarizes very little!

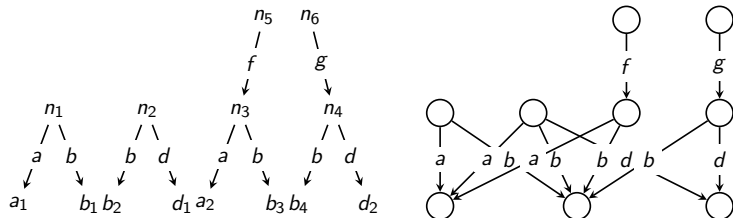
**Solution:** Bounded bisimilarity, e.g.,  $\sim_{1fb}$  (the weakest)

# Common graph quotients: bisimilarity [HHK95]

Two nodes are forward (resp. backward) bisimilar if they have exactly the same incoming (resp. outgoing) paths;  $\sim_{fw}$ ,  $\sim_{bw}$ ,  $\sim_{fb}$

**Problem:** Bisimilarity compresses/summarizes very little!

**Solution:** Bounded bisimilarity, e.g.,  $\sim_{1fb}$  (the weakest)



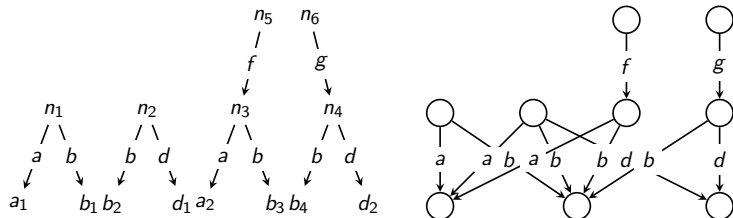
Still:  $> 130$  property combinations on conf. papers in DBLP

# Common graph quotients: bisimilarity [HHK95]

Two nodes are forward (resp. backward) bisimilar if they have exactly the same incoming (resp. outgoing) paths;  $\sim_{fw}$ ,  $\sim_{bw}$ ,  $\sim_{fb}$

**Problem:** Bisimilarity compresses/summarizes very little!

**Solution:** Bounded bisimilarity, e.g.,  $\sim_{1fb}$  (the weakest)



Still:  $> 130$  property combinations on conf. papers in DBLP

**Requirement 1:** We need equivalence relationships robust to structural heterogeneity.

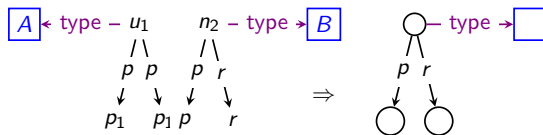
# What about type and schema triples?

Can we apply quotientization directly to an RDF graph?



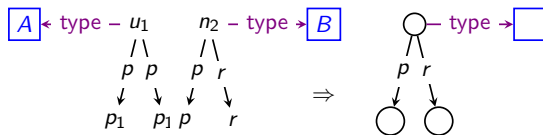
# What about type and schema triples?

Can we apply quotientization directly to an RDF graph?  
 Sample graph  $G$  and a possible quotient:



# What about type and schema triples?

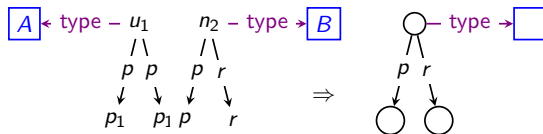
Can we apply quotientization directly to an RDF graph?  
 Sample graph  $G$  and a possible quotient:



Possible loss of class and property names

# What about type and schema triples?

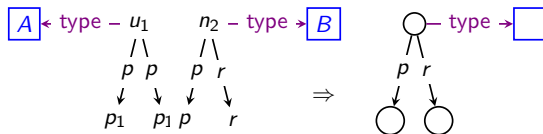
Can we apply quotientization directly to an RDF graph?  
 Sample graph G and a possible quotient:



Possible loss of schema triples

# What about type and schema triples?

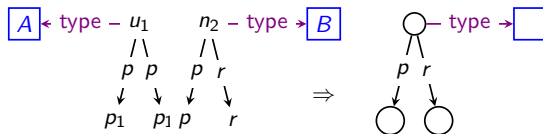
Can we apply quotientization directly to an RDF graph?  
 Sample graph  $G$  and a possible quotient thereof:



Possible loss of implicit triples

# What about type and schema triples?

Can we apply quotientization directly to an RDF graph?  
 Sample graph  $G$  and a possible quotient thereof:



Possible loss of implicit triples

**Requirement 2:** Avoid schema loss through quotientization

# RDF equivalence relation and RDF summaries

To meet Requirement 2, we define:

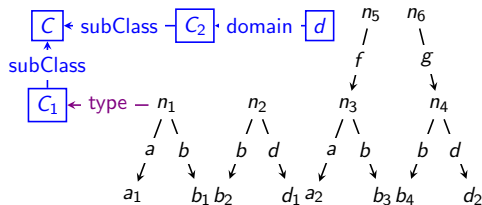
- 1 **RDF equivalence relation**: an equivalence relation on RDF graph nodes such that any class or property node is only equivalent to itself
- 2 **RDF summary**: a quotient of a graph  $G$  by an RDF equivalence relation such that any class or property node is represented by itself.

**Consequence:** For any RDF equivalence relation  $\equiv$  and RDF graph  $G$ , the schema of  $G_{/\equiv}$  is the schema of  $G$ .

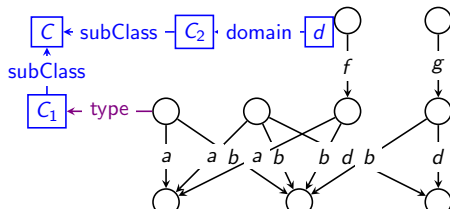
$\Rightarrow$  No schema compression! (to be rediscussed briefly)

# Summarization through an RDF equivalence relation

E.g., let  $\equiv_{1fb}$  to be the RDF node equivalence obtained from  $\sim_{1fb}$ .  
Sample graph G:



Its quotient through the RDF node equivalence  $\equiv_{1fb}$ :



# Recap

We have seen:

- **RDF node equivalence** and **RDF quotients**  $\Rightarrow$  structural representativeness, empty query pruning

We still need to solve:

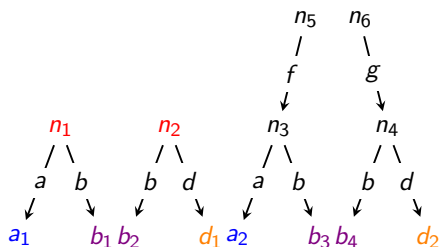
- **Requirement 1:** compact (yet preserve structure) even on heterogeneous graphs
- **Requirement 3:** can we summarize implicit triples?

We will address them in this order.



# RDF node equivalence based on property cliques

Intuition:  $n_1, n_2$  are “of the same kind”; similarly  $b_1, b_2, b_3$

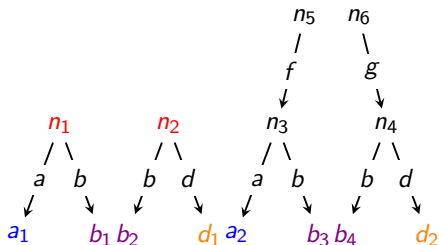


$n_3, n_4$  may or may not be of the same kind as  $n_1, n_2$ .

# RDF node equivalence based on property cliques

**Output property cliques:**  $\{a, b, d\}$ ;  $\{f\}$ ;  $\{g\}$ ;  $\emptyset$

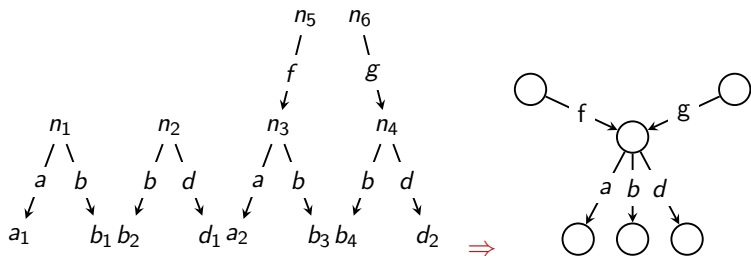
**Input property cliques:**  $\{a\}$ ;  $\{b\}$ ;  $\{d\}$ ;  $\{f\}$ ;  $\{g\}$ ;  $\emptyset$



## Weak clique-based summaries

Two nodes are weakly equivalent ( $\equiv_{/W}$ ) iff they have **the same input clique** **or** **the same output clique** **or** are weakly equivalent to a third one.

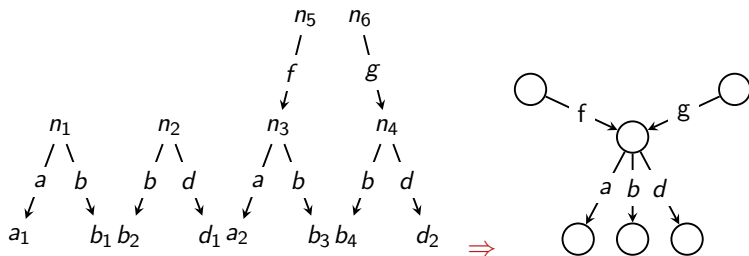
Weak summary  $G_{/W}$  of the sample RDF graph  $G$ :



## Weak clique-based summaries

Two nodes are weakly equivalent ( $\equiv_{/W}$ ) iff they have **the same input clique** **or** **the same output clique** **or** are weakly equivalent to a third one.

Weak summary  $G_{/W}$  of the sample RDF graph  $G$ :

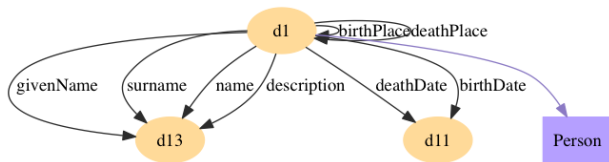


**Property:** In  $G_{/W}$ , each data property appears exactly once  $\Rightarrow$  its nodes are “source of  $p$ , target of  $p$ ” for each  $p$  [ČGM15b].

## Weak clique-based summaries

**Property:**  $G/W$  nodes are “source of  $p$ , target of  $p$ ” for each  $p$ .

**Detecting errors in the data:** why do the birthplace and deathplace loop?



Looking in the data, we find:

---

```

<http://dbpedia.org/resource/Kunitomo_Ikkansai> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person> .

```

---

```

<http://dbpedia.org/resource/Kunitomo_Ikkansai> <http://dbpedia.org/ontology/birthPlace>
<http://dbpedia.org/resource/Kunitomo_Ikkansai> .

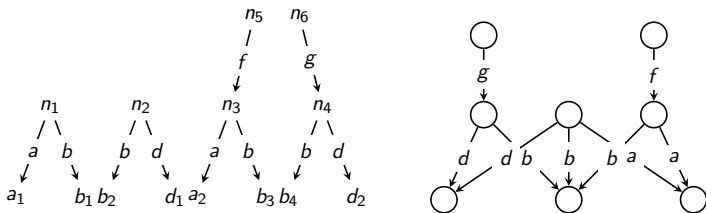
```

---

## Strong clique-based summaries

Two nodes are strongly equivalent ( $\equiv_S$ ) iff they have **the same input clique** **and** **the same output clique**.

Strong summary  $G_{/\equiv_S}$  of the same  $G$ :



## Which role should node types play in summarization?

Having the same type(s) is orthogonal w.r.t. having the same structure.

## Which role should node types play in summarization?

Having the same type(s) is orthogonal w.r.t. having the same structure. Two alternatives:

- 1 **Data-then-type:** group nodes first by their data triples, then carry the types from each  $\equiv$  group to its representative.

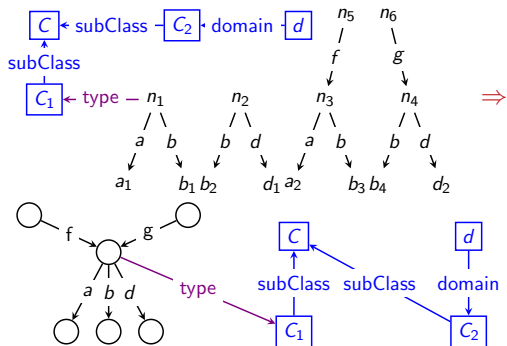


# Which role should node types play in summarization?

Having the same type(s) is orthogonal w.r.t. having the same structure. Two alternatives:

- 1 **Data-then-type:** group nodes first by their data triples, then carry the types from each  $\equiv$  group to its representative.

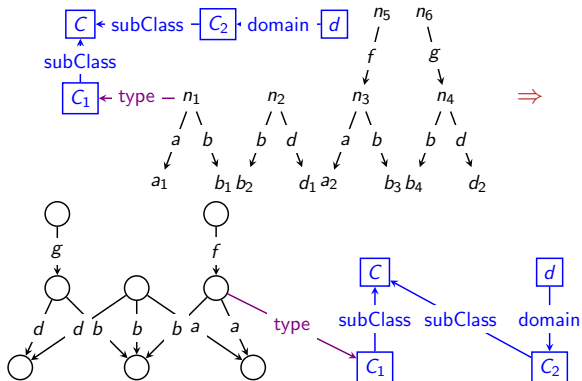
Extended Weak summary:



# Adding types after data summarization

- Data-then-type:** group nodes first by their data triples, then carry the types from each  $\equiv$  group to its representative.

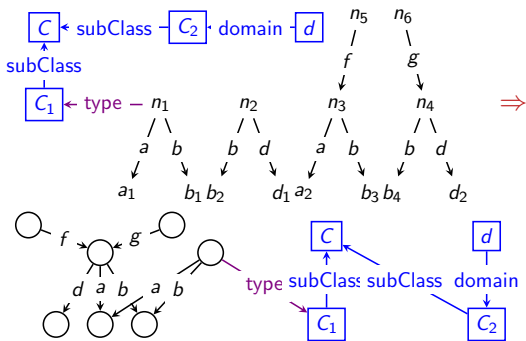
Extended Strong summary:



## Giving prominence to types

- 2 **Type-then-data:** Group nodes by their type set, and **untyped** nodes by their data properties.

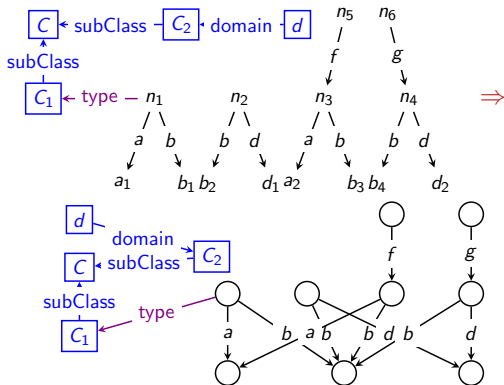
Typed Weak summary  $G_{/\equiv TW}$  of the sample graph:



# Giving prominence to types

- Type-then-data:** Group nodes first by their types. Only untyped nodes are grouped by their data properties.

Typed Strong summary  $G_{/\equiv}^{TS}$  of the sample graph:



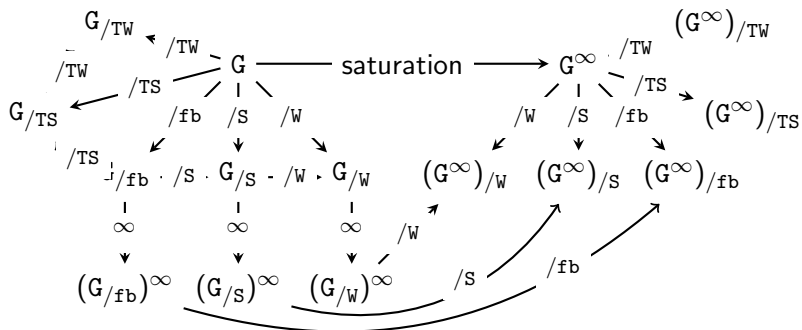
# RDF summaries outline

| Summary       | Weak? | Strong? | Types first? |
|---------------|-------|---------|--------------|
| $G \equiv W$  | ✓     |         |              |
| $G \equiv S$  |       | ✓       |              |
| $G \equiv TW$ | ✓     |         | ✓            |
| $G \equiv TS$ |       | ✓       | ✓            |

# RDF summaries outline

| Summary          | Weak? | Strong? | FW bisim? | BW bisim? | Types first? |
|------------------|-------|---------|-----------|-----------|--------------|
| $G \equiv W$     | ✓     |         |           |           |              |
| $G \equiv S$     |       | ✓       |           |           |              |
| $G \equiv TW$    | ✓     |         |           |           | ✓            |
| $G \equiv TS$    |       | ✓       |           |           | ✓            |
| $G \equiv fw$    |       |         | ✓         |           |              |
| $G \equiv bw$    |       |         |           | ✓         |              |
| $G \equiv fb$    |       |         | ✓         | ✓         |              |
| $G \equiv fw, T$ |       |         | ✓         |           | ✓            |
| $G \equiv bw, T$ |       |         |           | ✓         | ✓            |
| $G \equiv fb, T$ |       |         | ✓         | ✓         | ✓            |

## Relations between RDF summaries [ČGM17b]



## Summary size comparison (more in [ČGM17b])

| Graph G | G           | Summary $G_{/\equiv}$ | $ G_{/\equiv} $   | $cf_{\equiv}$ |
|---------|-------------|-----------------------|-------------------|---------------|
| DBLP    | 150,787,464 | $G_{/W}$              | <b>71</b>         | 2,123,767     |
| DBLP    | 150,787,464 | $G_{/S}$              | <b>206</b>        | 731,978       |
| DBLP    | 150,787,464 | $G_{/fw}$             | <b>262,695</b>    | 574           |
| LUBM1M  | 1,227,868   | $G_{/W}$              | <b>161</b>        | 7,579         |
| LUBM1M  | 1,227,868   | $G_{/S}$              | <b>207</b>        | 5,903         |
| LUBM1M  | 1,227,868   | $G_{/fw}$             | <b>1982</b>       | 617           |
| LUBM10M | 11,990,183  | $G_{/W}$              | <b>162</b>        | 74,013        |
| LUBM10M | 11,990,183  | $G_{/S}$              | <b>206</b>        | 58,204        |
| LUBM10M | 11,990,183  | $G_{/fw}$             | <b>24,958</b>     | 480           |
| LUBM10M | 11,990,183  | $G_{/bw}$             | <b>6,162</b>      | 1,944         |
| LUBM10M | 11,990,183  | $G_{/fb}$             | <b>11,990,076</b> | 1             |



# Summarizing $G^\infty$

Recall: With an RDF Schema, the semantics of  $G$  is  $G^\infty \Rightarrow$

We really need  $(G^\infty)_{/\equiv}$ !

- 1 Saturate  $G$ , then summarize
- 2 Can we avoid saturating  $G$ ?...

# Summarizing $G^\infty$

Recall: With an RDF Schema, the semantics of  $G$  is  $G^\infty \Rightarrow$   
 We really need  $(G^\infty)_{/\equiv}$ !

- ① Saturate  $G$ , then summarize
- ② Can we avoid saturating  $G$ ?...

## Shortcut theorem [ČGM17a]

For the summaries  $G_{/\equiv} W$ ,  $G_{/\equiv} S$ ,  $G_{/\equiv} fw$ ,  $G_{/\equiv} bw$ ,  $G_{/\equiv} fb$ :

$(G^\infty)_{/\equiv}$  is the same as  $((G_{/\equiv})^\infty)_{/\equiv}$

Also: **sufficient condition** for any  $\equiv$  to admit the shortcut.

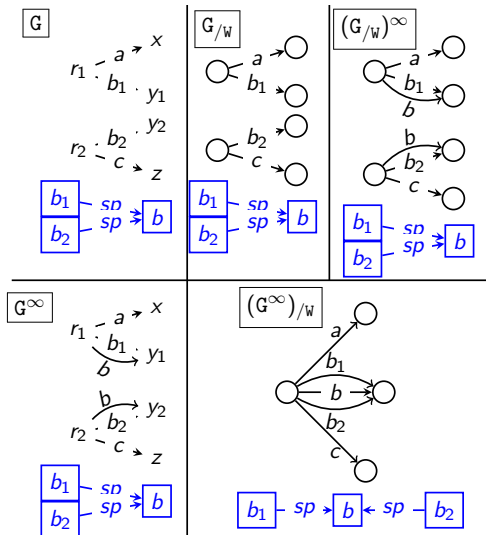
# Shortcut path to $G^\infty$

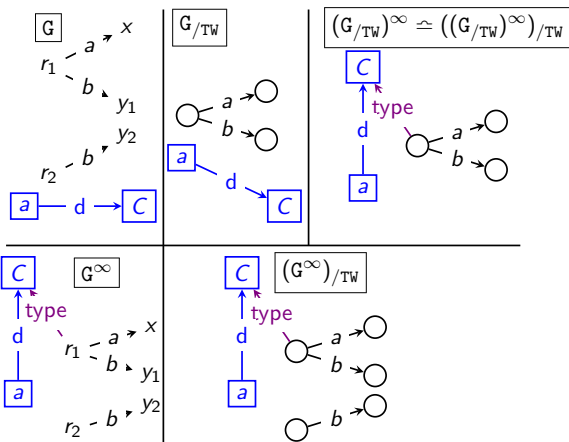
Direct  $G \rightarrow \mathbf{sat.} \rightarrow G^\infty \rightarrow \mathbf{summ.} \rightarrow (G^\infty)_\equiv$

Shortcut  $G \rightarrow \mathbf{summ.} \rightarrow G_\equiv \rightarrow \mathbf{sat.} \rightarrow (G_\equiv)^\infty \rightarrow \mathbf{summ.} \rightarrow ((G_\equiv)^\infty)_\equiv$

If  $G_\equiv$  is much smaller than  $G$ , **the shortcut may be faster!**

Up to 20 times in our experiments [ČGM17b]

Shortcut example:  $G/\equiv W$ 

Shortcut counter-example:  $G/\equiv TW$ 

# Summarization algorithms

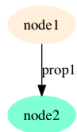
- 1 **Global algorithms:** visit all  $G$ , compute  $\equiv$  relation, then traverse  $G$  again and represent each triple in  $G_{/\equiv} W$
- 2 **Incremental algorithms:** visit  $G$ , compute  $\equiv$  and summary based on knowledge gained so far; **adjust** summary

We devised global and incremental summarization algorithms for  $G_{/\equiv} W$ ,  $G_{/\equiv} S$ ,  $G_{/\equiv} TW$ ,  $G_{/\equiv} TS$ .

Difficulty of incremental summarization: adjusting  $\equiv$  and revisiting summarization decisions

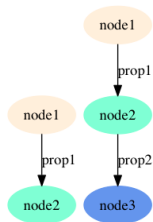
## Example: weak incremental summarization (1)

Each color corresponds to a different  $\equiv_W$  class



## Example: weak incremental summarization (1)

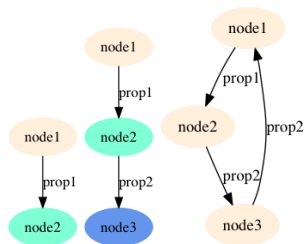
Each color corresponds to a different  $\equiv_W$  class





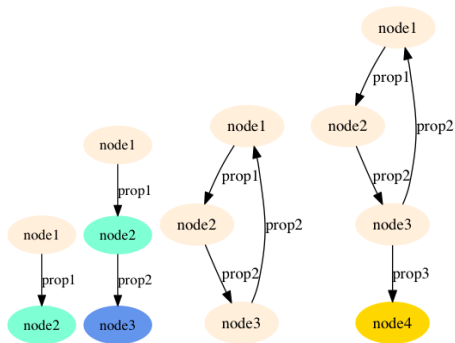
# Example: weak incremental summarization (1)

Each color corresponds to a different  $\equiv_W$  class



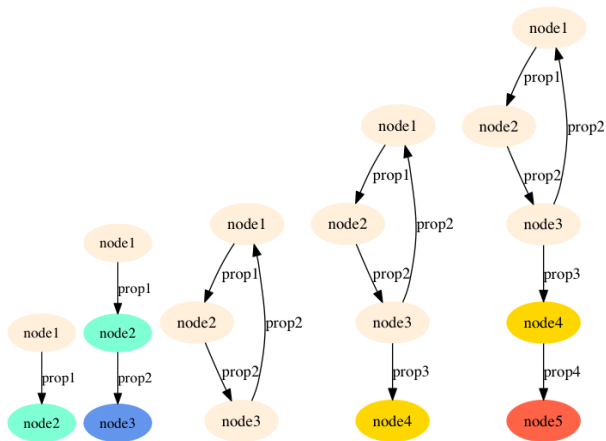
# Example: weak incremental summarization (1)

Each color corresponds to a different  $\equiv_W$  class



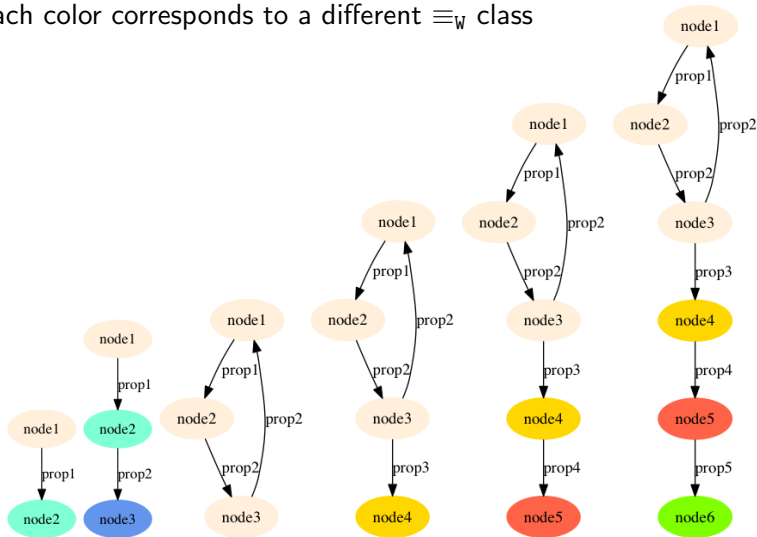
# Example: weak incremental summarization (1)

Each color corresponds to a different  $\equiv_W$  class



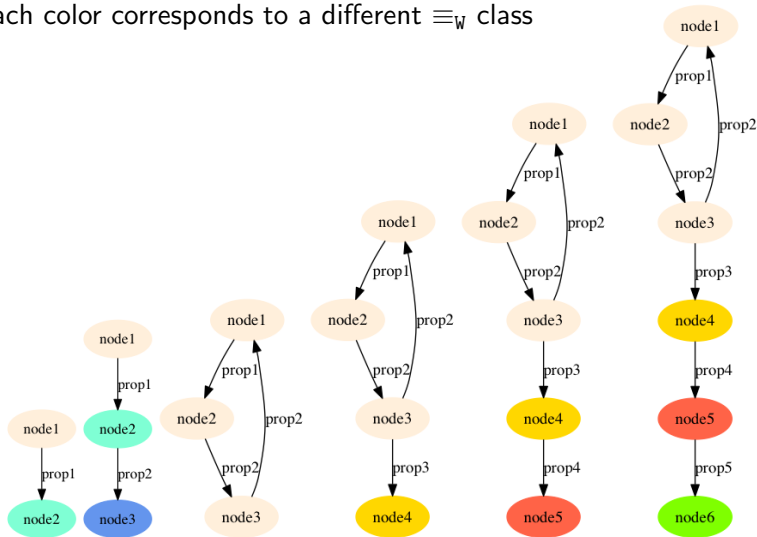
# Example: weak incremental summarization (1)

Each color corresponds to a different  $\equiv_W$  class

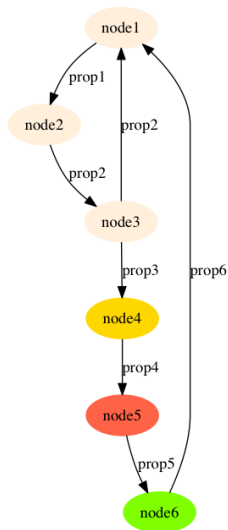


# Example: weak incremental summarization (1)

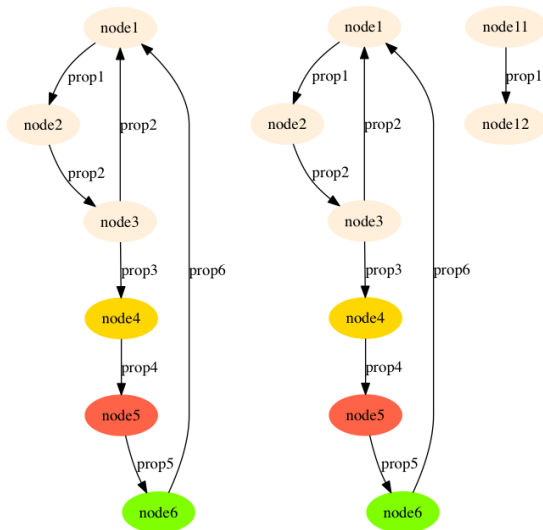
Each color corresponds to a different  $\equiv_W$  class



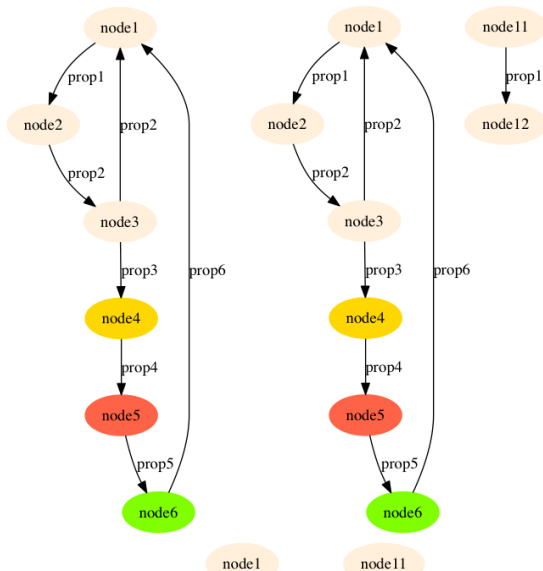
## Example: weak incremental summarization (2)



# Example: weak incremental summarization (2)

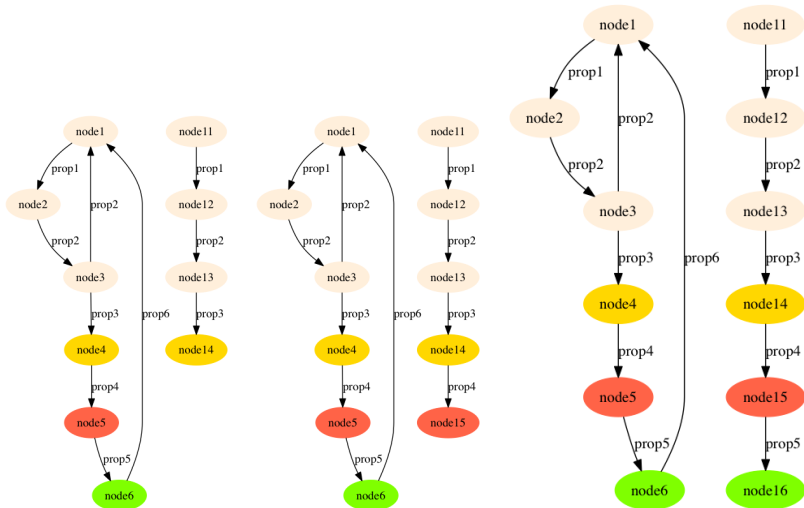


# Example: weak incremental summarization (2)



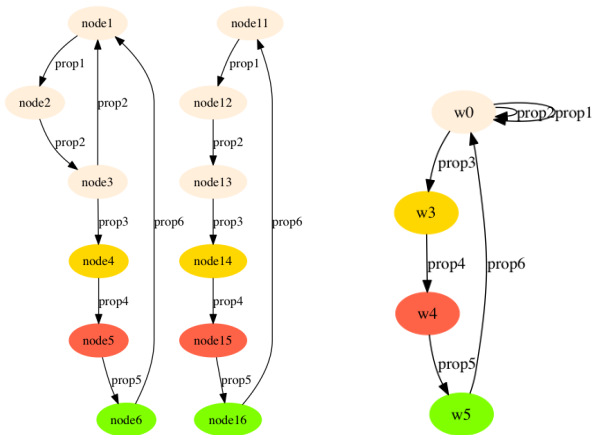


# Example: weak incremental summarization (3)



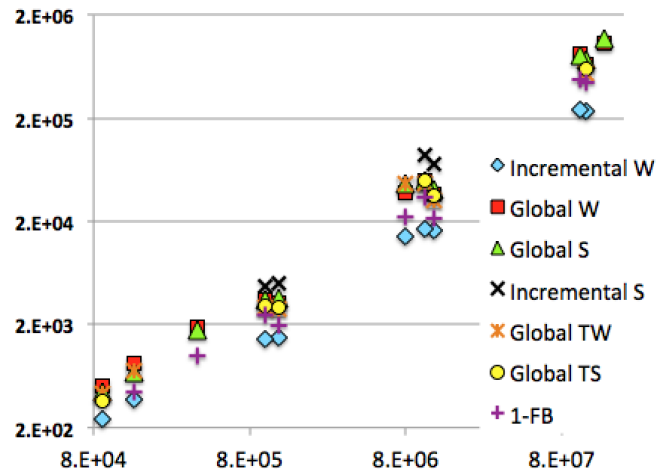
# Example: weak incremental summarization (end)

Full graph and its summary:



# Algorithm scale-up

$10^5$  to  $1.5 \times 10^8$  triples




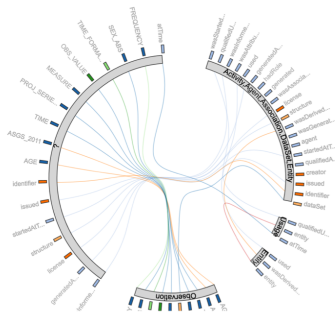
# Summary-enabled LOD cloud exploration [PGA<sup>+</sup>18]

Collaboration with ILDA Inria data visualization team on LODAtlas

<http://lodatlas.lri.fr/>

Use summary to derive visualisation instead of the original graph  
(smaller, faster)

abs-linked-data : Australian Bureau of Statistics (ABS) Linked Data 



# Part IV

## Conclusion

# The need for RDF graph discovery tools

- RDF graphs can be **large and complex**, they lack a prescriptive schema
- Semantic rules lead to **implicit data**
- **Structural quotient summaries** compactly represent graph structure and semantics.  
Available online at: (new version soon)

<https://team.inria.fr/cedar/projects/rdfsummary/>

- Type-first summarization variant to cope with large type hierarchies [GM18]
- Integration into LODAtlas platform [PGA<sup>+</sup>18]

# Part V

## Perspectives

# Ongoing and future work

## Ongoing:

- 1 Experiment with new, parallel summarization algorithms based on Spark
- 2 Keyword search in RDF graphs based on quotient summaries

## Future:

- 1 Controlled inclusion of data value synopsis in the summary
- 2 Extension to more expressive ontology languages
- 3 Integration in a larger platform for summary-based data discovery (with Mirjana Mazuran)
- 4 Exploration of interesting aggregate view of RDF graphs (with Yanlei Diao)



# References

- [CDT13] Stéphane Campinas, Renaud Delbru, and Giovanni Tummarello. Efficiency and precision trade-offs in graph summary algorithms. In IDEAS, 2013.
- [CFKP15] Mariano P. Consens, Valeria Fionda, Shahan Khatchadourian, and Giuseppe Pirrò. S+EPPs: Construct and explore bisimulation summaries + optimize navigational queries (demo). PVLDB, 8(12), 2015.
- [ČGK<sup>+</sup>18] Šejla Čebirić, François Goasdoué, Haridimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika. Summarizing semantic graphs: A survey. 2018.
- [ČGM15a] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. Query-oriented summarization of RDF graphs. In BICOD, 2015.
- [ČGM15b] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. Query-oriented summarization of RDF graphs (demonstration). PVLDB, 8(12), 2015.
- [ČGM17a] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. A framework for efficient representative summarization of RDF graphs. In International Semantic Web Conference (ISWC), 2017.
- [ČGM17b] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. Query-Oriented Summarization of RDF Graphs. Research Report RR-8920, INRIA, 2017.

## References (cont.)

- [DMS17] Yanlei Diao, Ioana Manolescu, and Shu Shang. Dagger: Digging for interesting aggregates in RDF graphs. In International Semantic Web Conference (ISWC), 2017.
- [GM18] Paweł Guzewicz and Ioana Manolescu. Quotient RDF summaries based on type hierarchies. In Data Engineering for the Semantic Web (DESWeb), 2018.
- [GW97] Roy Goldman and Jennifer Widom. Dataguides: Enabling query formulation and optimization in semistructured databases. In VLDB, 1997.
- [HHK95] Monika Rauch Henzinger, Thomas A. Henzinger, and Peter W. Kopke. Computing simulations on finite and infinite graphs. In FOCS, 1995.
- [LYL13] Shou-De Lin, Mi-Yen Yeh, and Cheng-Te Li. Sampling and summarization for social networks (tutorial), 2013.
- [NM11] Thomas Neumann and Guido Moerkotte. Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins. In ICDE, 2011.
- [PGA<sup>+</sup>18] Emmanuel Pietriga, Hande Gözükan, Caroline Appert, Marie Destandau, Šejla Čebirić, François Goasdoué, and Ioana Manolescu. Browsing linked data catalogs with LODAtlas. In Int'l. Semantic Web Conference (ISWC), Resources track, 2018.

## References (cont.)

- [ZDYZ14] Haiwei Zhang, Yuanyuan Duan, Xiaojie Yuan, and Ying Zhang. ASSG: adaptive structural summary for RDF graph data. In ISWC (Posters and Demonstrations), 2014.