Centre for
Data Analytics

Insight

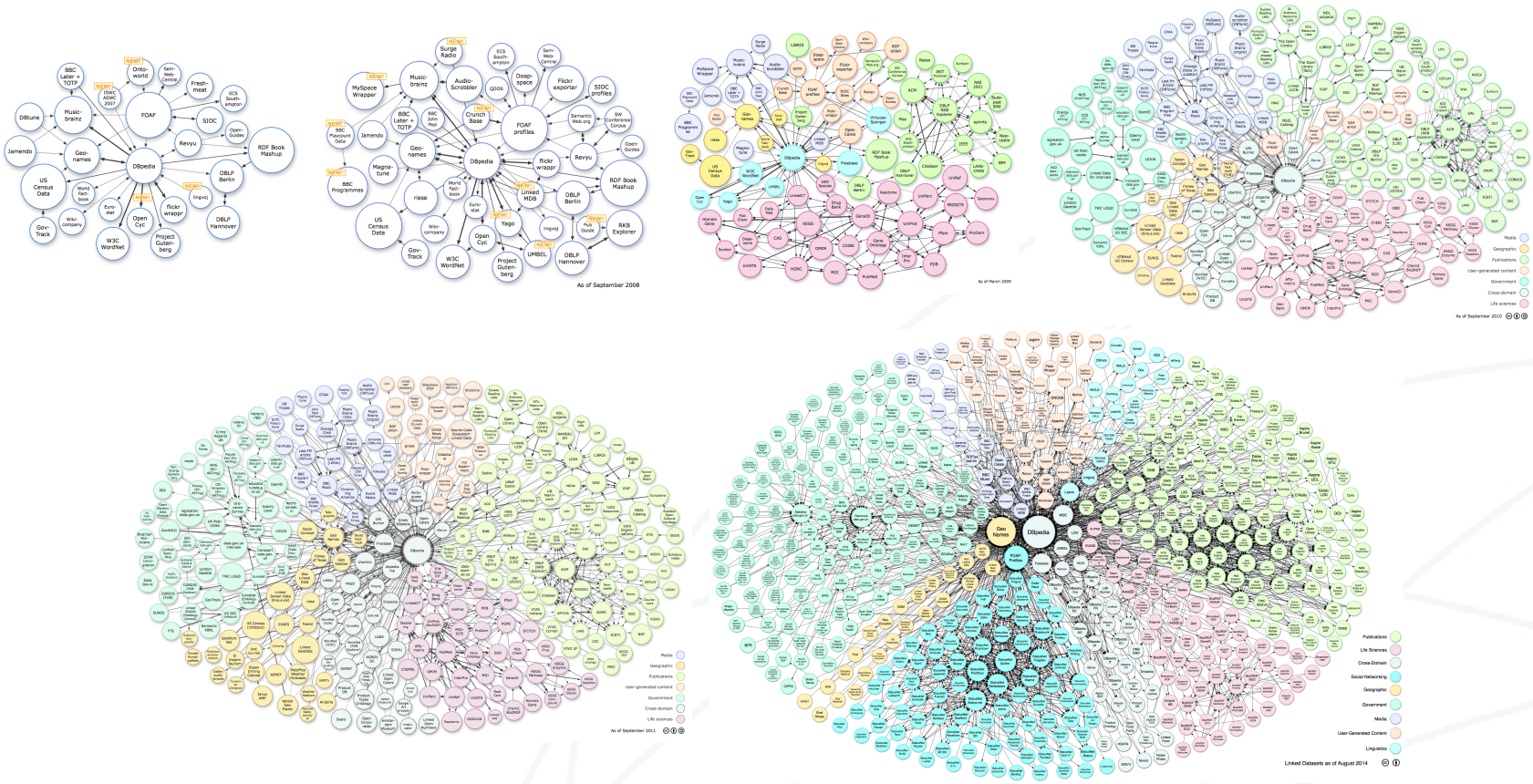
Making Sense of Web Data

Stéphane Campinas

Thesis VIVA, 13 December 2016

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN SFI Science Foundation Ireland

Evolution of Web Data



Linking Open Data cloud diagrams 2008-2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. `http://lod-cloud.net/`

Creation of a Schema-like Data Structure

Top-Down Schema: Ontologies

- Data has a known structure. . .
- . . . but it does not keep up with the pace of Web Data evolution

Bottom-Up Schema: Graph Summarisation

- Embraces the Semantic Web vision. . .
- . . . but is sensitive to data quality

Creation of a Schema-like Data Structure

Top-Down Schema: Ontologies

- Data has a known structure. . .
- . . . but it does not keep up with the pace of Web Data evolution

Bottom-Up Schema: Graph Summarisation

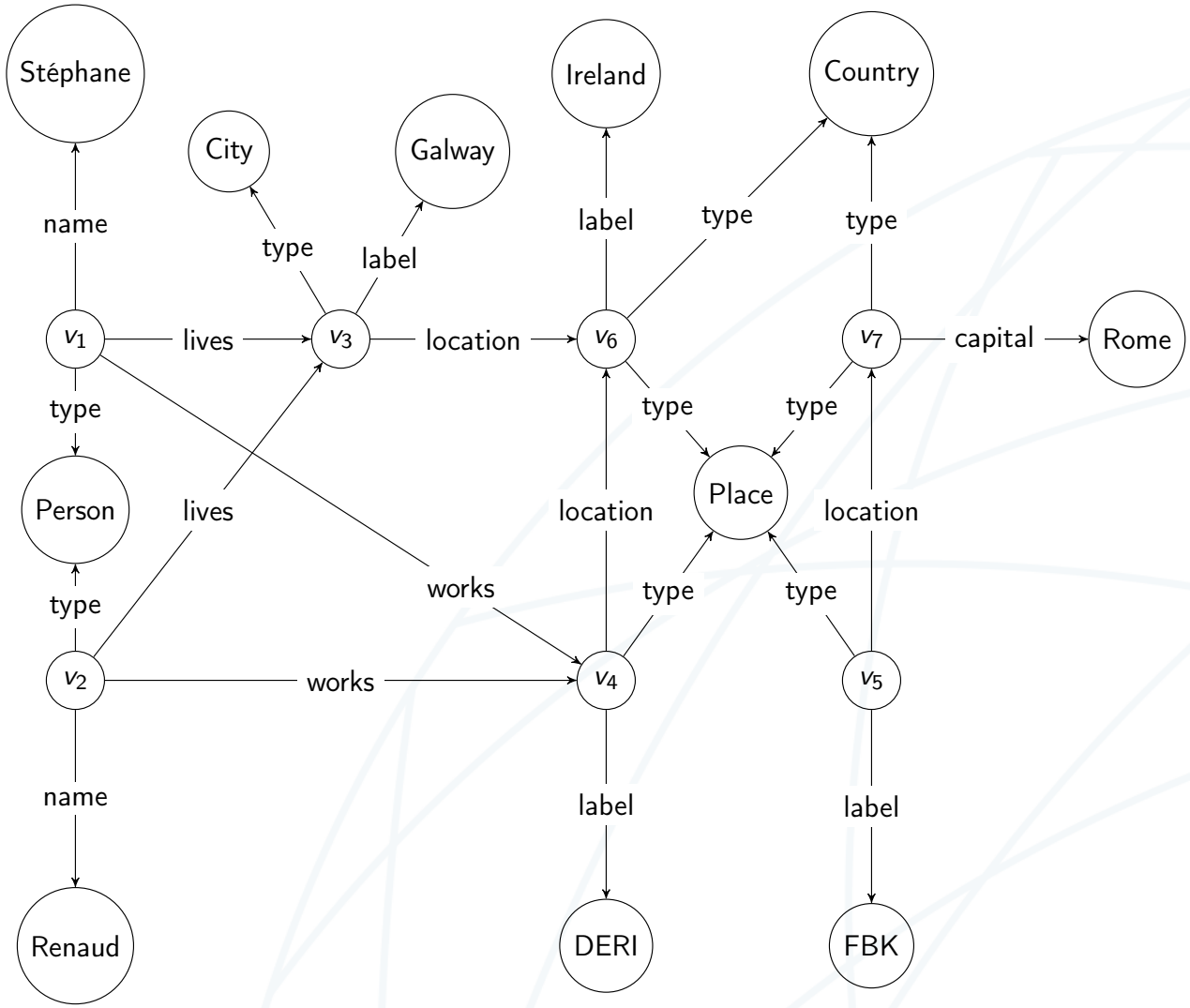
- Embraces the Semantic Web vision. . .
- . . . but is sensitive to data quality

Challenges

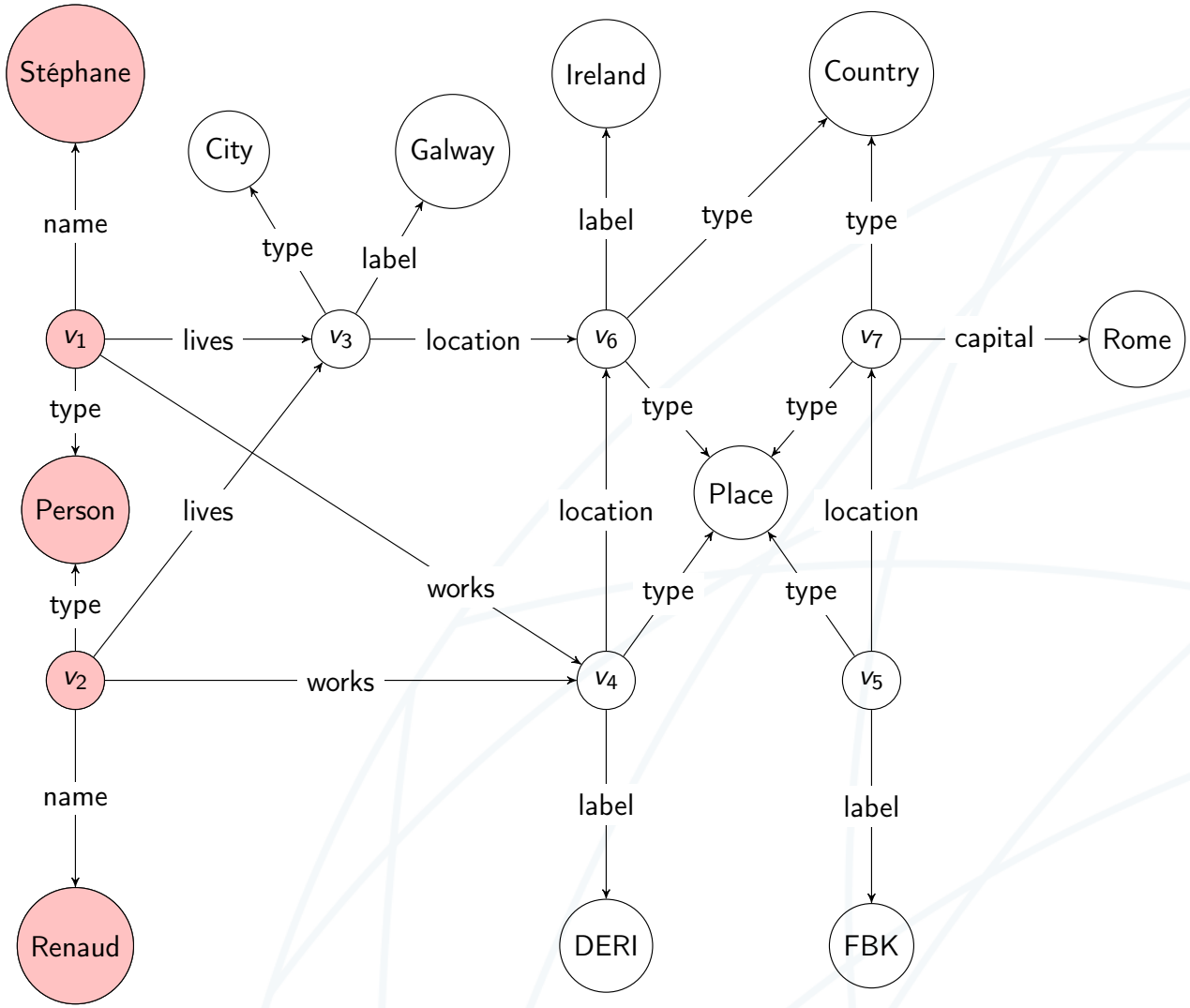
Graph Summarisation

- How to generate a description of a dataset ?
- Scale to billions of triples
- How to measure the precision of the generated output ?

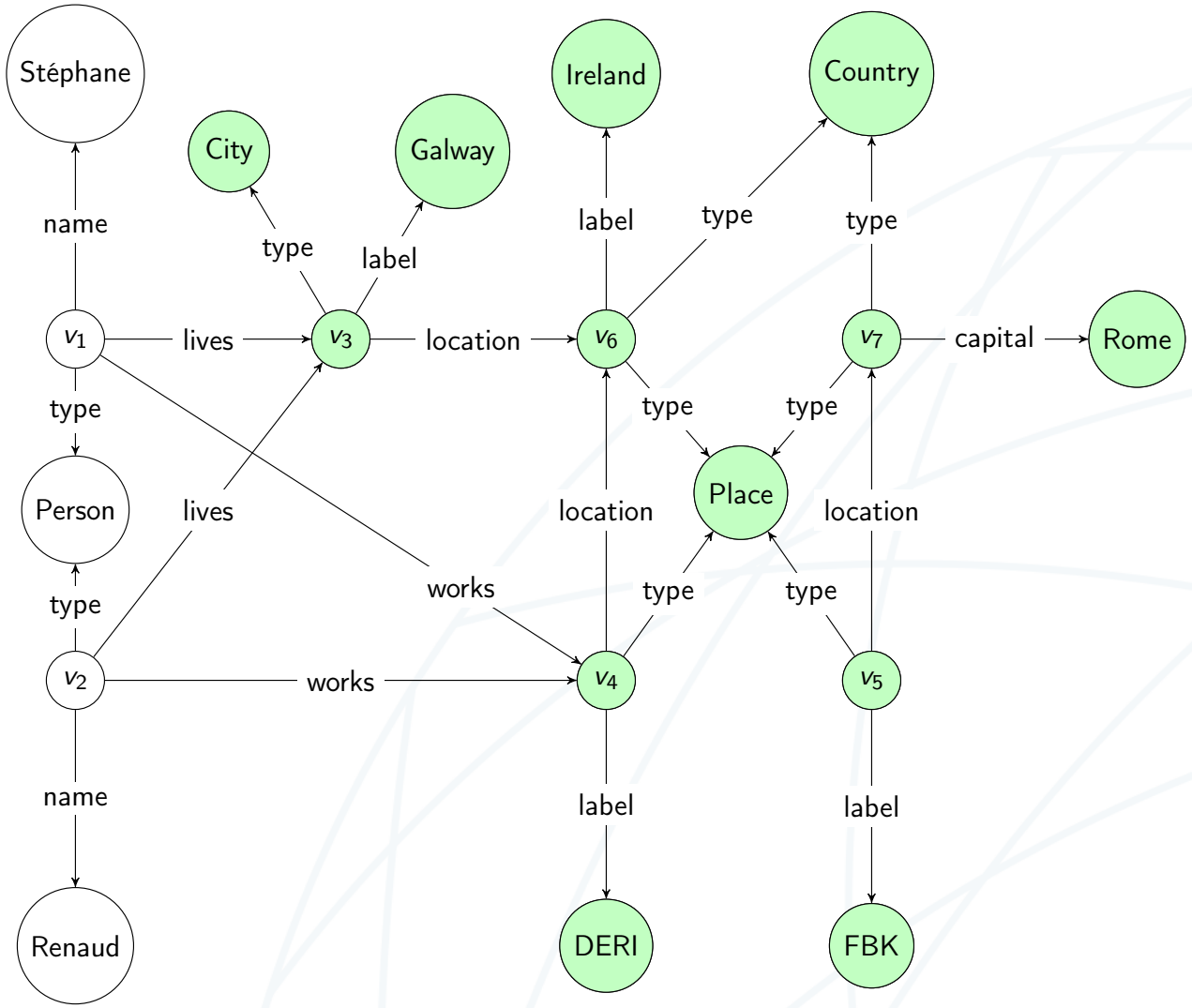
Graph Data Model



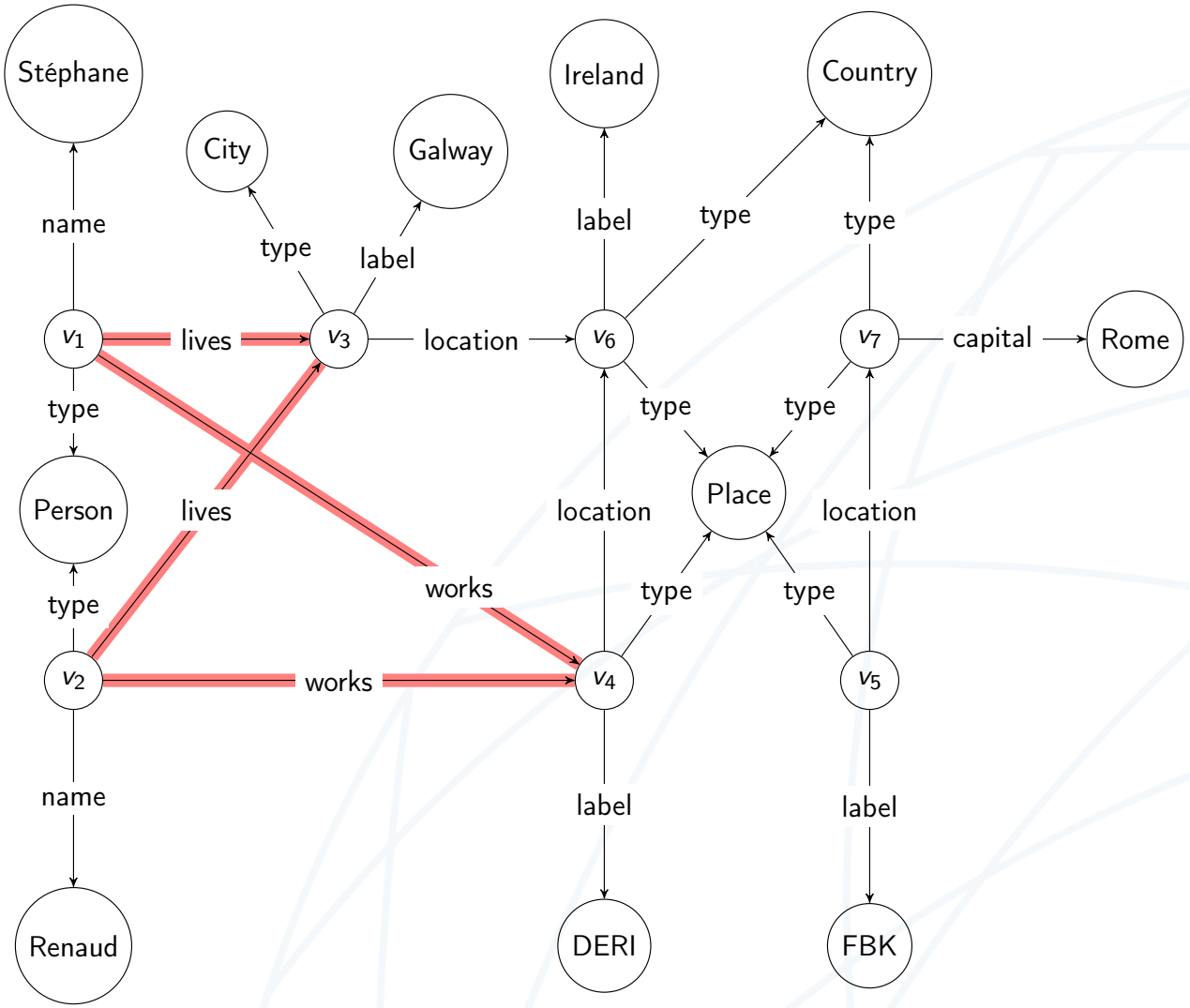
Graph Data Model



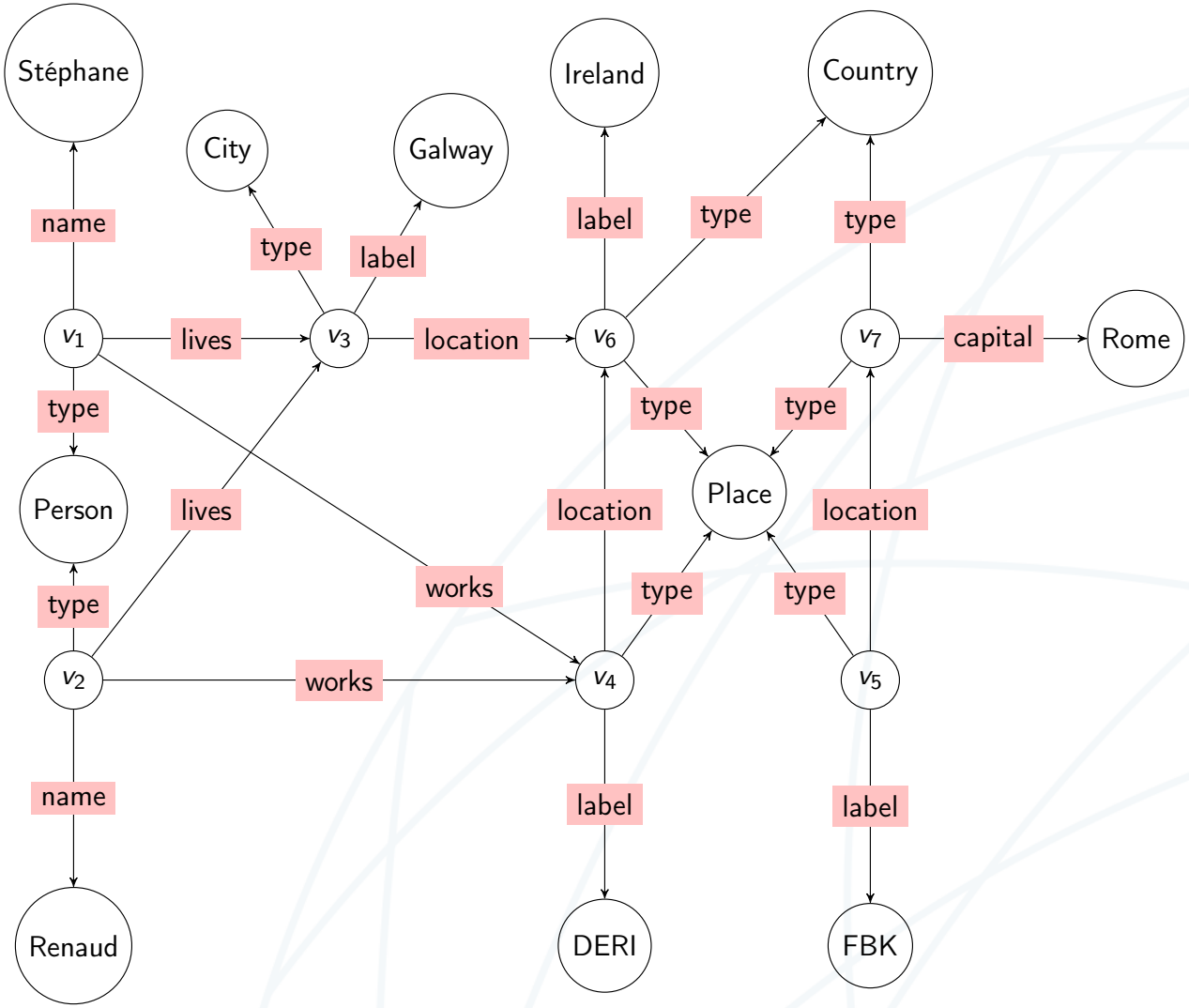
Graph Data Model



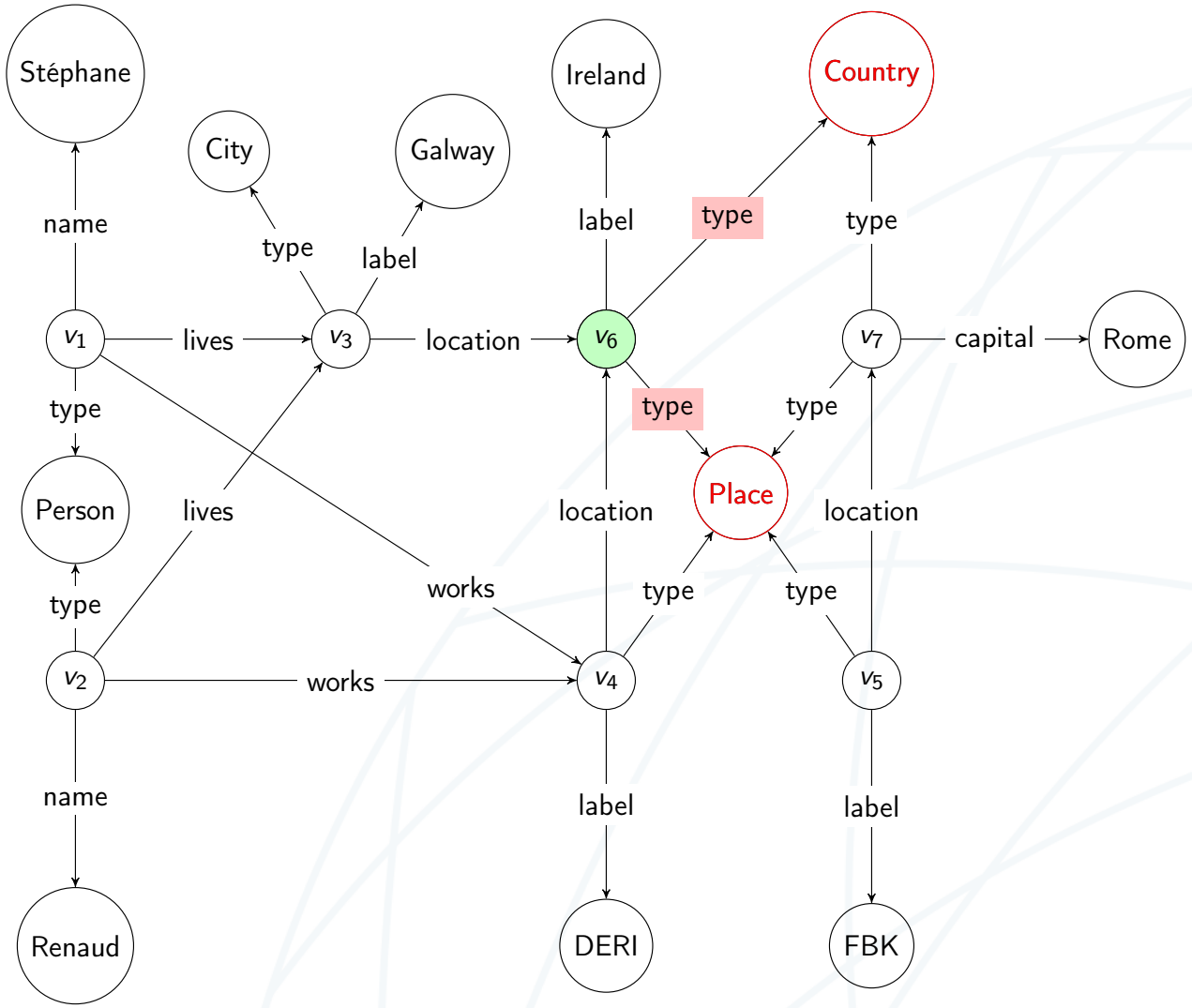
Graph Data Model



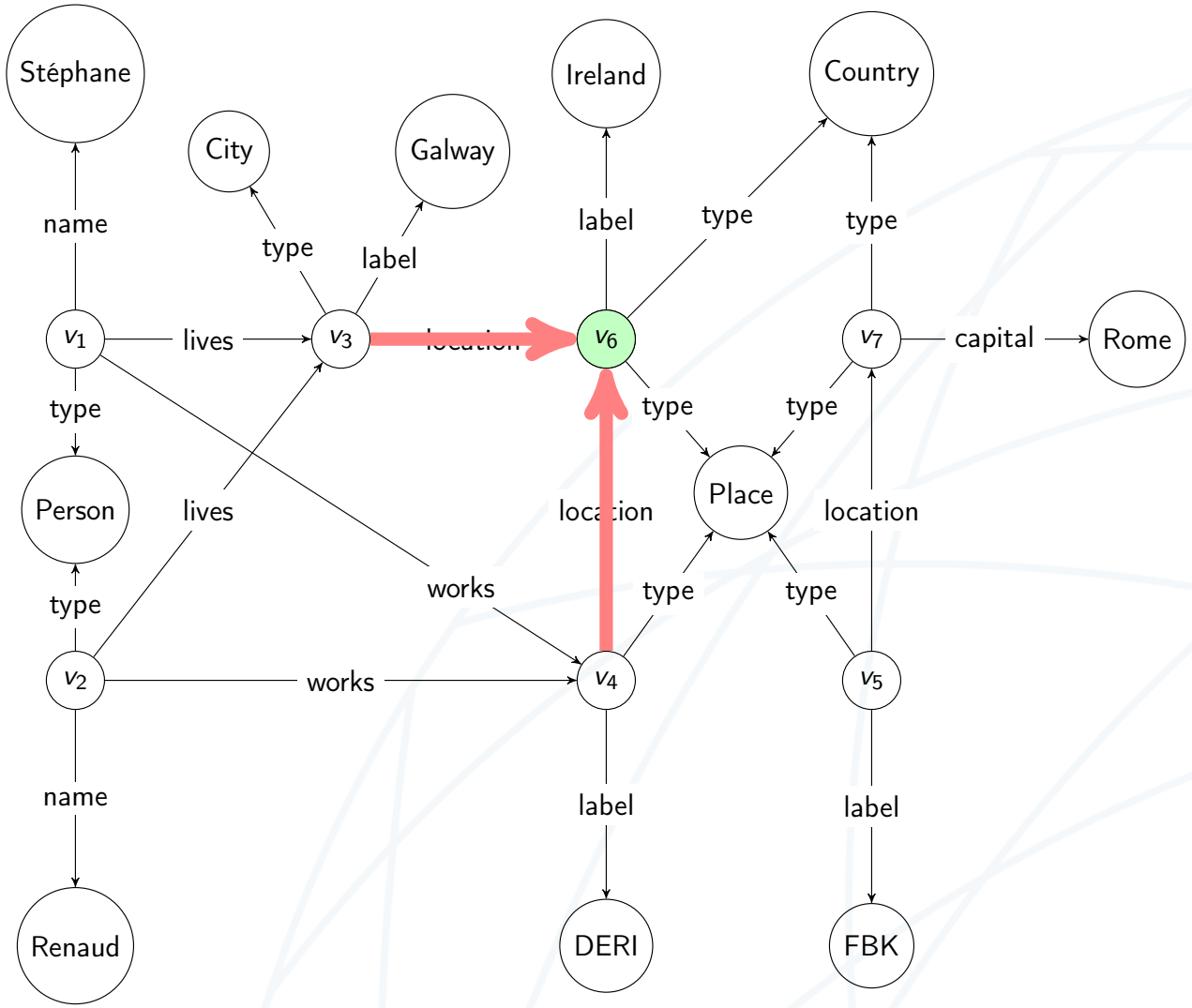
Graph Data Model



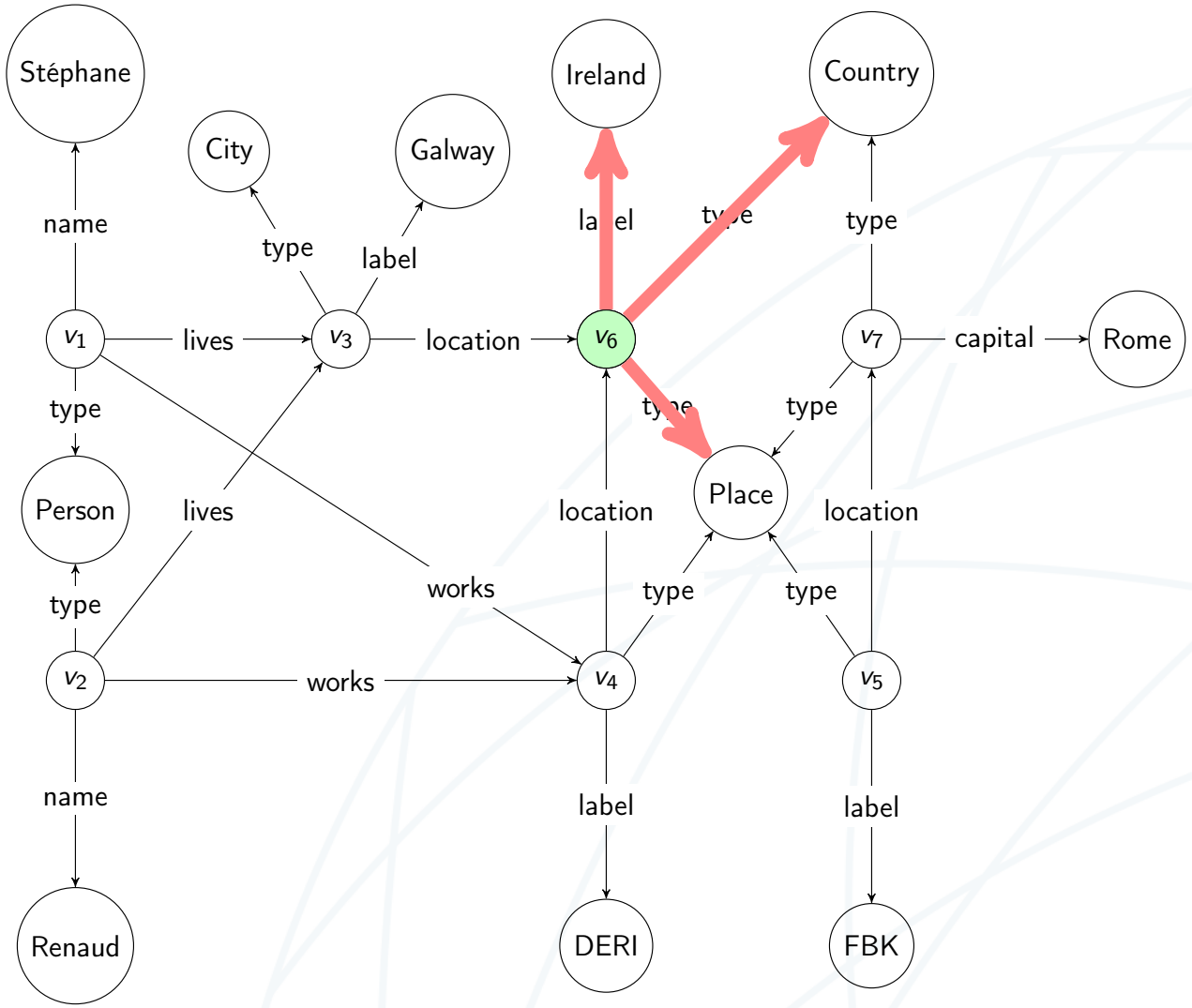
Graph Data Model



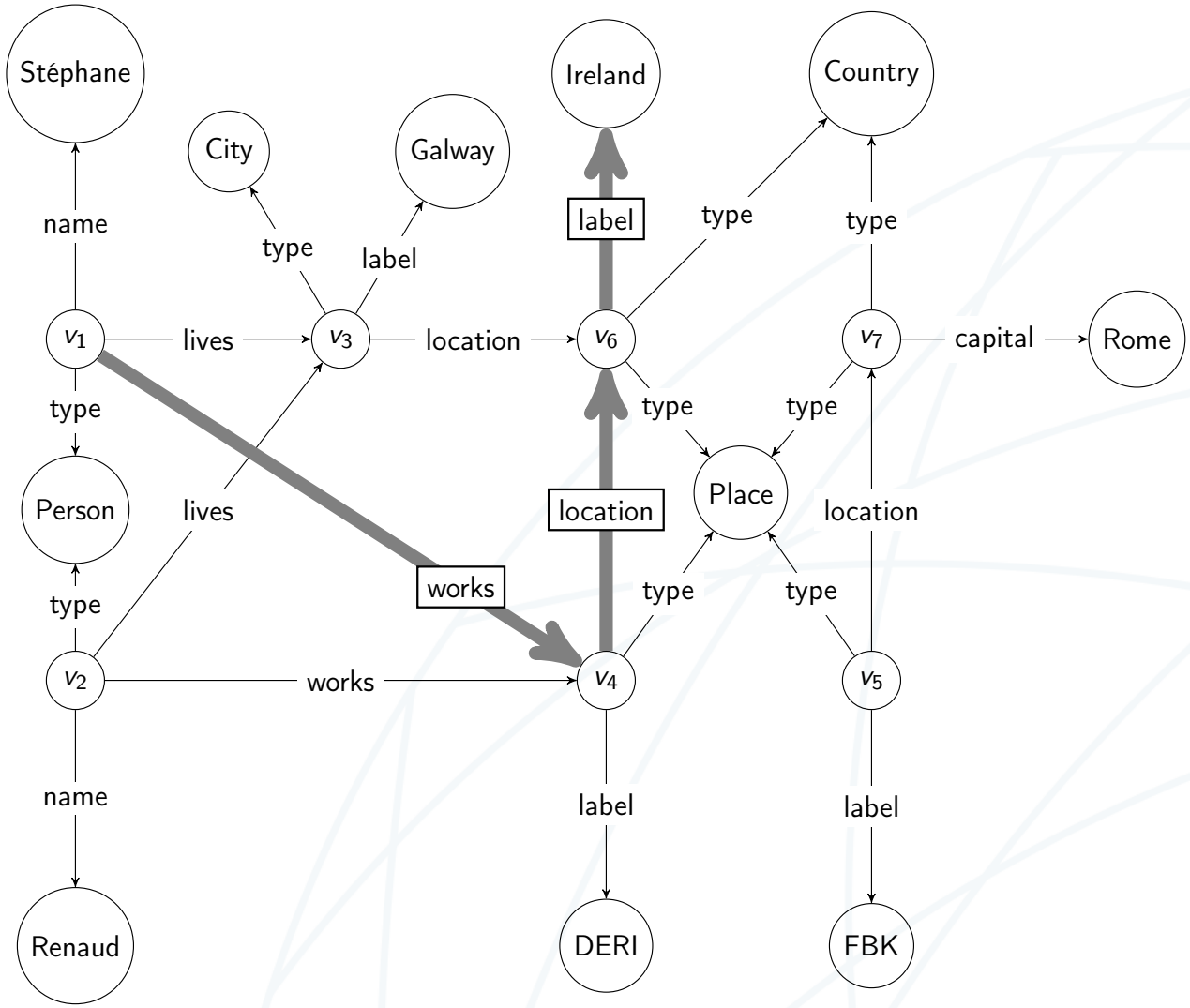
Graph Data Model



Graph Data Model



Graph Data Model



Entity Model

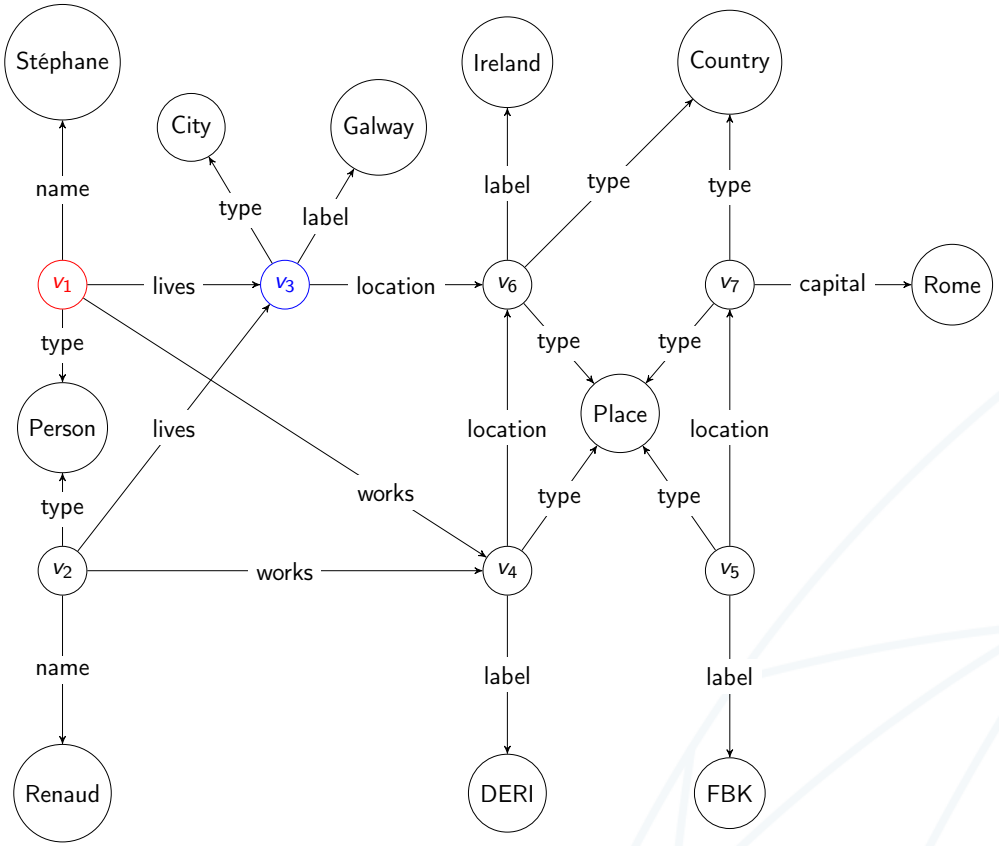


Figure: An entity graph describing people, places, and their relationships.

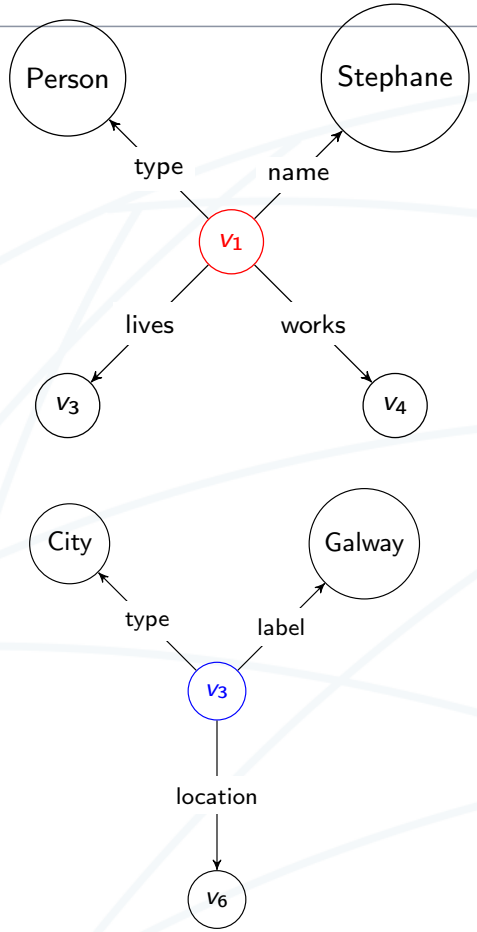
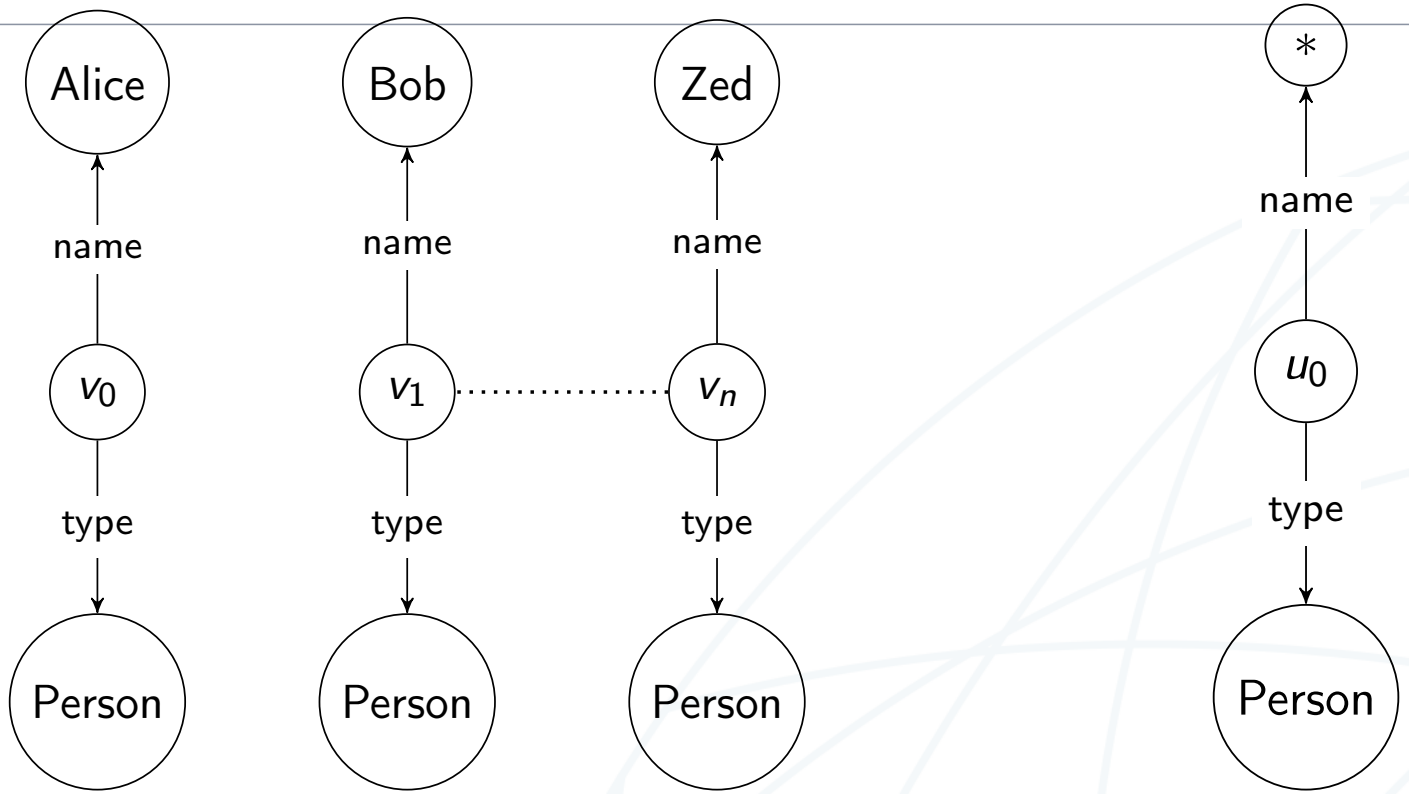


Figure: Entity Descriptions

Overview



(a) An entity graph about people

(b) A possible summary

Figure: Summarising an entity graph

Graph Summary

Definition

A graph is the **summary** of another graph if there is a **homomorphism** from the *latter* to the *former* with respect to a **relation**.

Summarisation Relation

- Mapping between the nodes of two graphs, based on *features* of the graph

Graph Homomorphism

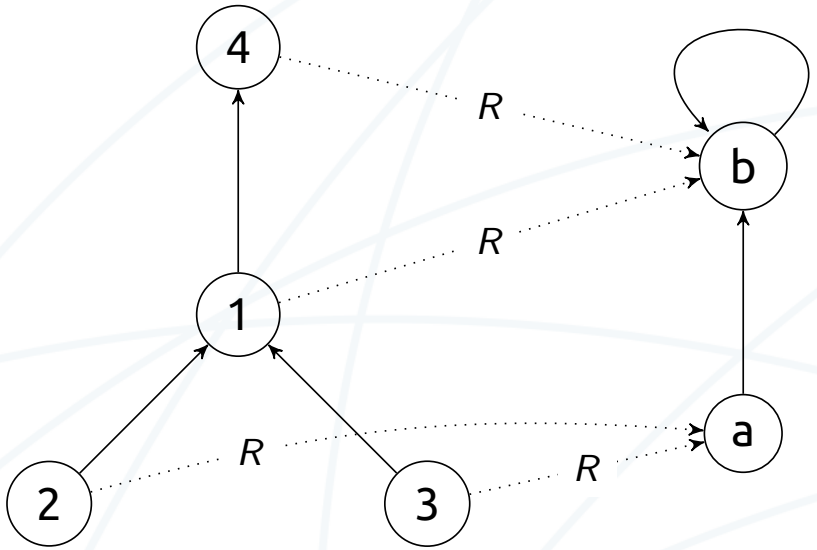
- With respect to the summarisation relation, each edge of the graph is mapped to an edge(s) of the summary

Graph Homomorphism

Example

Definition

A graph is homomorphic to another if all of its edges do exist in that other graph for a given relation.

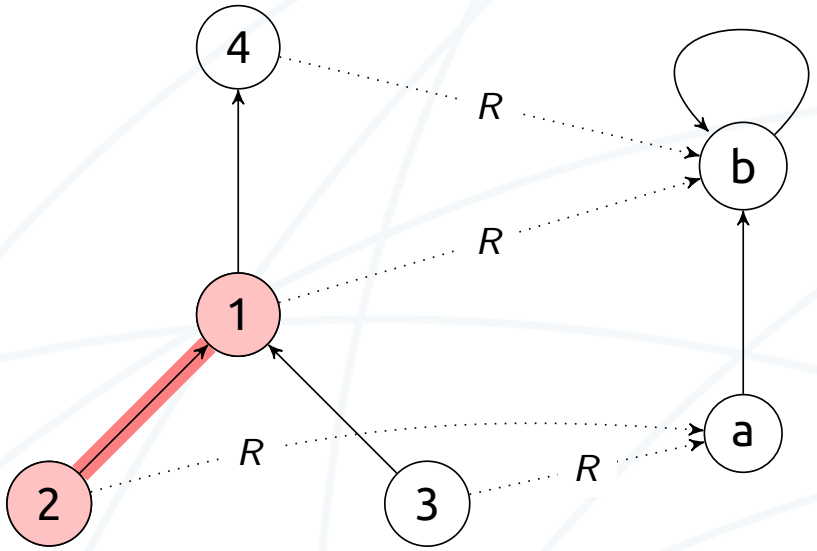


Graph Homomorphism

Example

Definition

A graph is homomorphic to another if all of its edges do exist in that other graph for a given relation.

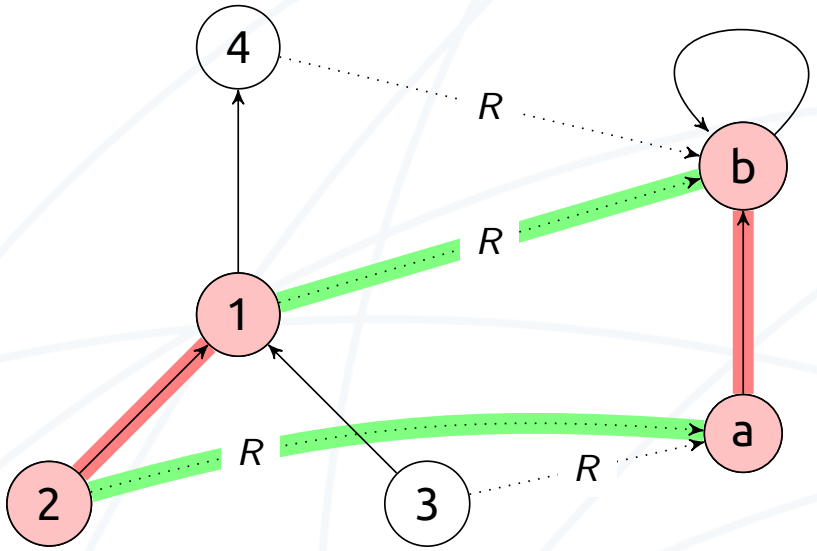


Graph Homomorphism

Example

Definition

A graph is homomorphic to another if all of its edges do exist in that other graph for a given relation.

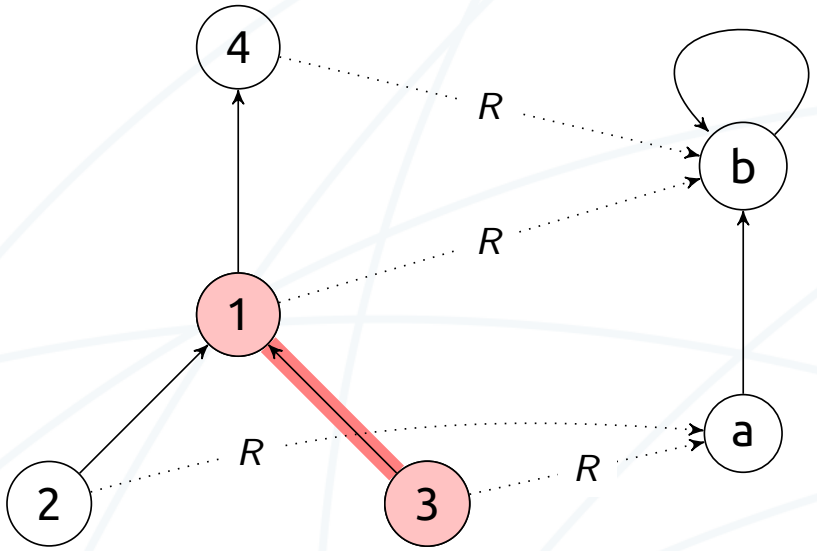


Graph Homomorphism

Example

Definition

A graph is homomorphic to another if all of its edges do exist in that other graph for a given relation.

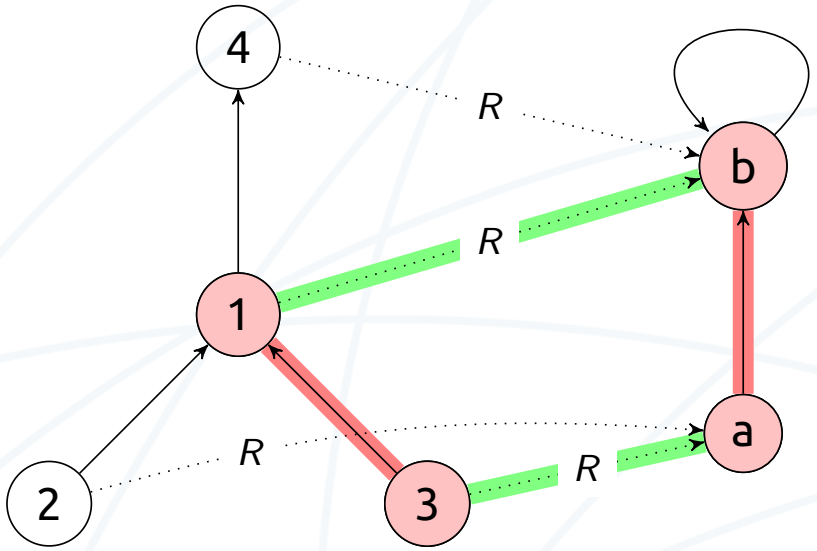


Graph Homomorphism

Example

Definition

A graph is homomorphic to another if all of its edges do exist in that other graph for a given relation.

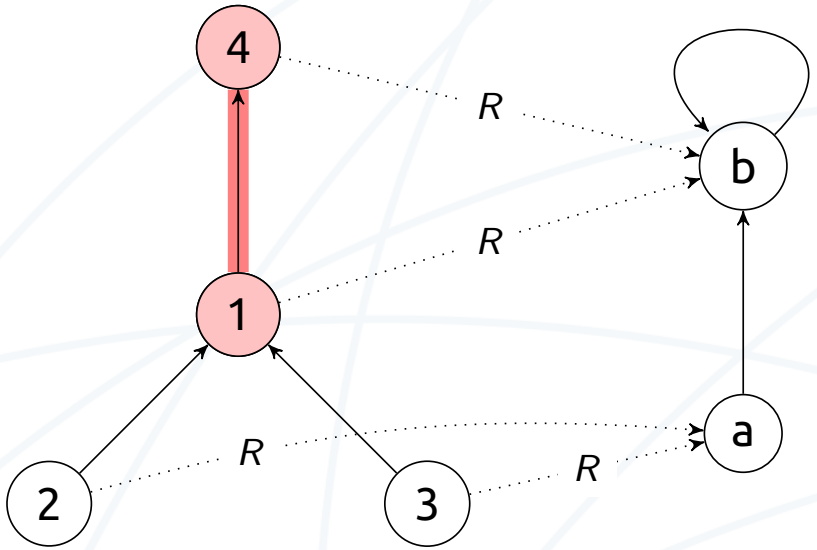


Graph Homomorphism

Example

Definition

A graph is homomorphic to another if all of its edges do exist in that other graph for a given relation.

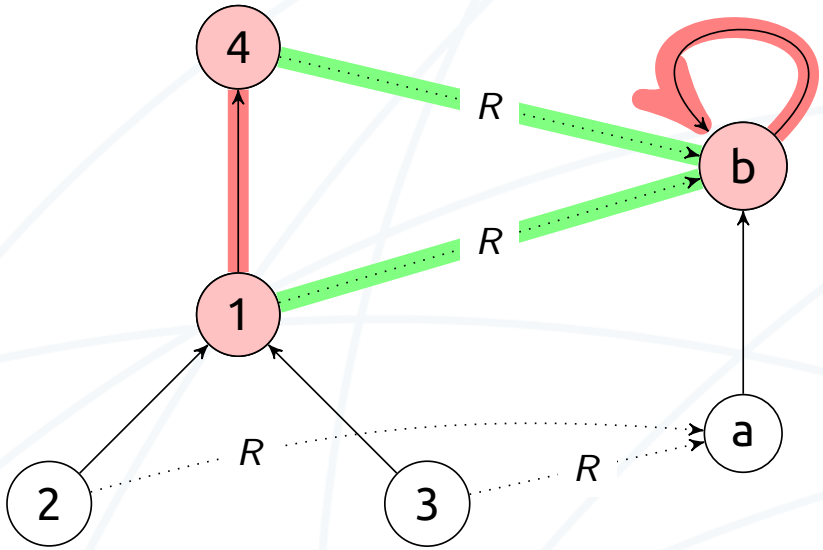


Graph Homomorphism

Example

Definition

A graph is homomorphic to another if all of its edges do exist in that other graph for a given relation.



Structural and Vocabulary Heterogeneity

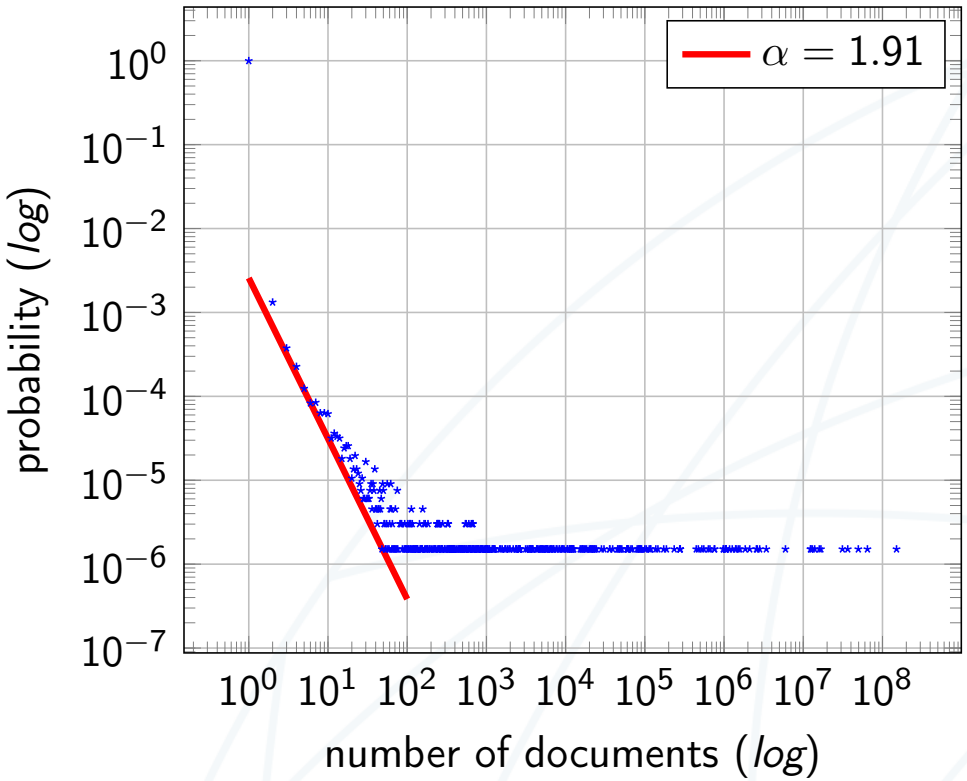


Figure: Ontology probability distribution

Approximate Graph Summaries

Precise Graph Summary: F&B-Bisimulation

- A summary is precise if every path in the summary do exist
- Creating such a summary is impractical due to the data's heterogeneity

Approximate Graph Summary

- Summarisation needs to scale to large graphs
- Computationally expensive

What is being approximated ? Paths existence.

Approximate Graph Summaries

Precise Graph Summary: F&B-Bisimulation

- A summary is precise if every path in the summary do exist
- Creating such a summary is impractical due to the data's heterogeneity

Approximate Graph Summary

- Summarisation needs to scale to large graphs
- Computationally expensive

What is being approximated ? Paths existence.

Generic Graph Summarisation Framework

Possible Summarisation Features

- Types
- Attributes (incoming and outgoing)

Studied Summaries

- Unique Type Summary
- Types Summary
- Attributes Summary
- Types & Attributes Summary
- IO Attributes Summary
- IO Attributes Types Summary

Generic Graph Summarisation Framework

Possible Summarisation Features

- Types
- Attributes (incoming and outgoing)

Studied Summaries

- Unique Type Summary
- Types Summary
- Attributes Summary
- Types & Attributes Summary
- IO Attributes Summary
- IO Attributes Types Summary

Example: Types Summary

Relation: set of types associated to a node

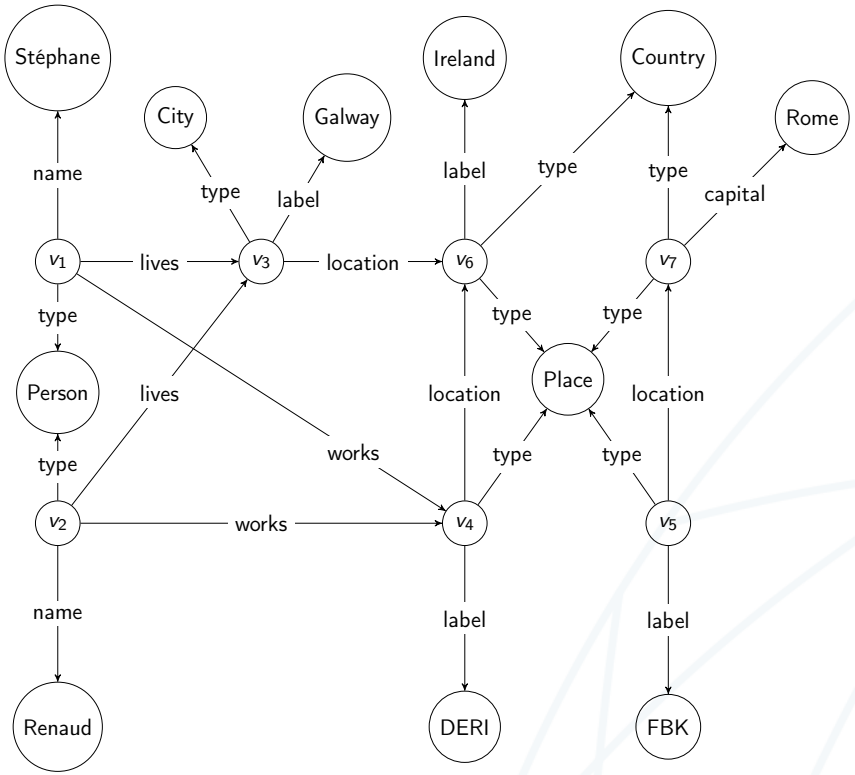


Figure: An entity graph describing people, places, and their relationships.

Figure: Types summary of the entity graph.

Example: Types Summary

Relation: set of types associated to a node

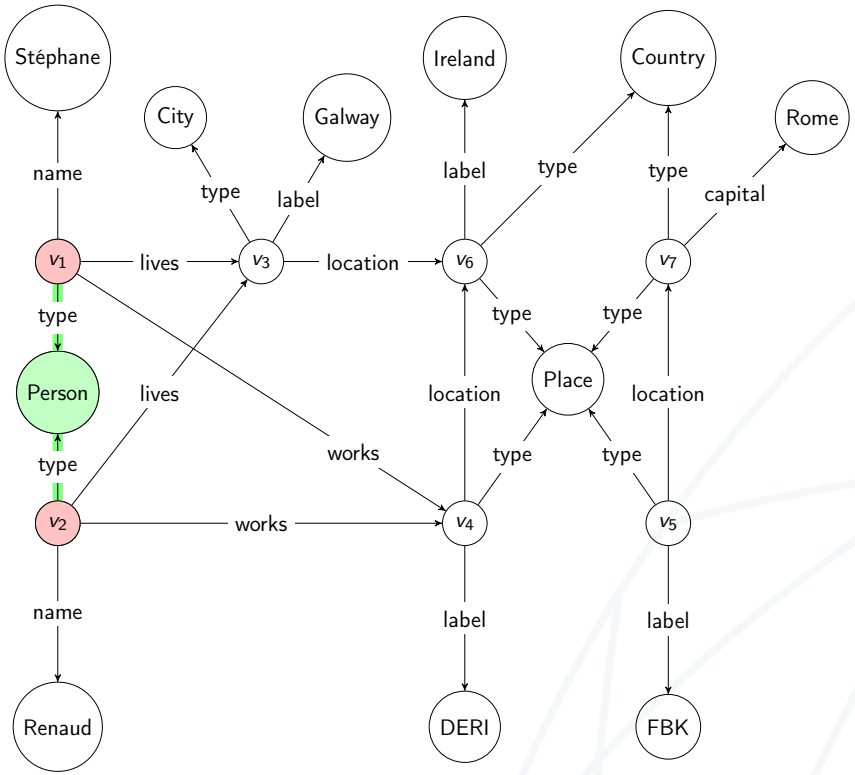


Figure: An entity graph describing people, places, and their relationships.

Figure: Types summary of the entity graph.

Example: Types Summary

Relation: set of types associated to a node

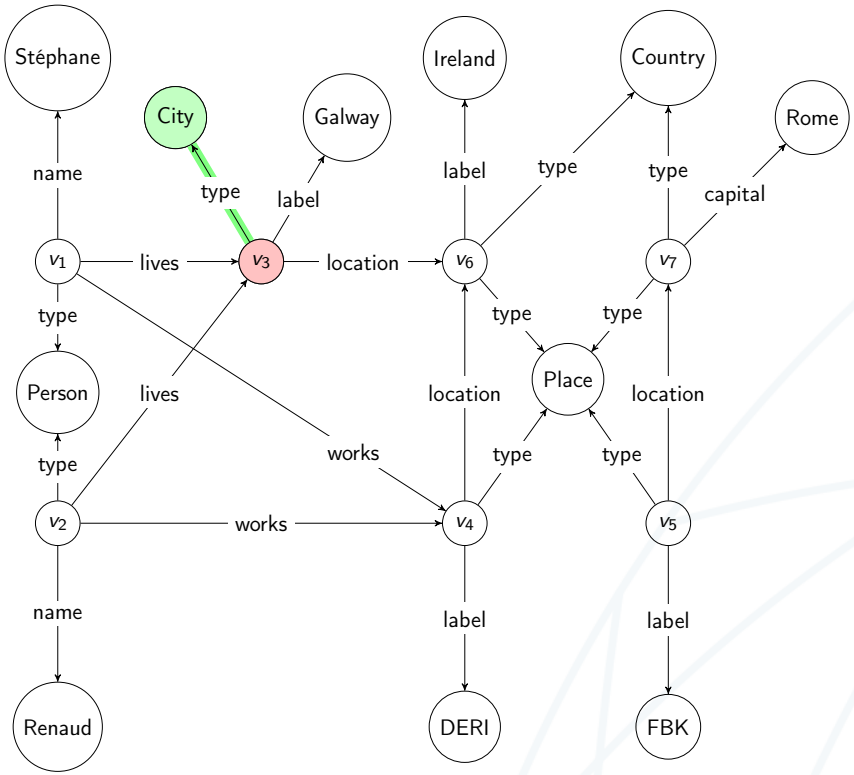


Figure: An entity graph describing people, places, and their relationships.

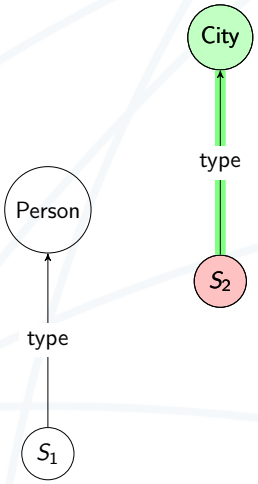


Figure: Types summary of the entity graph.

Example: Types Summary

Relation: set of types associated to a node

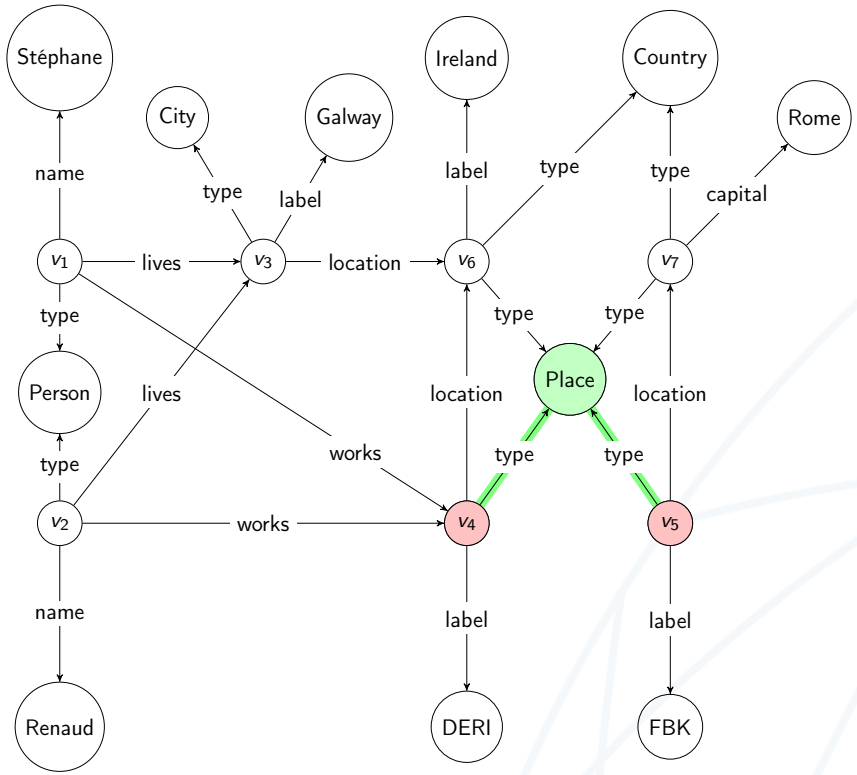


Figure: An entity graph describing people, places, and their relationships.

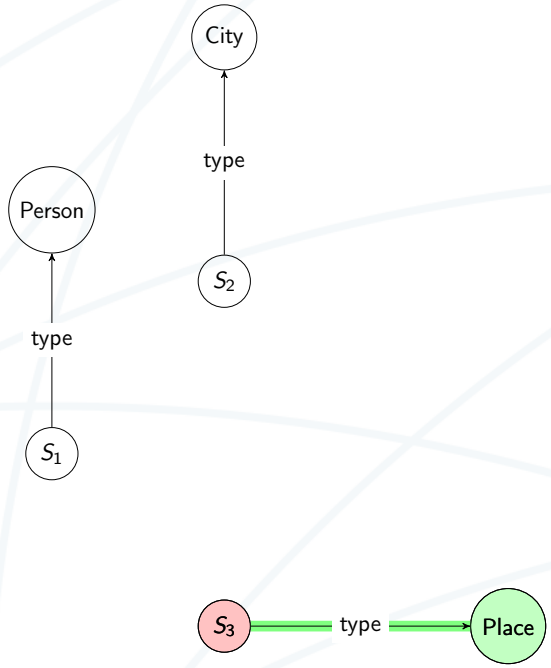


Figure: Types summary of the entity graph.

Example: Types Summary

Relation: set of types associated to a node

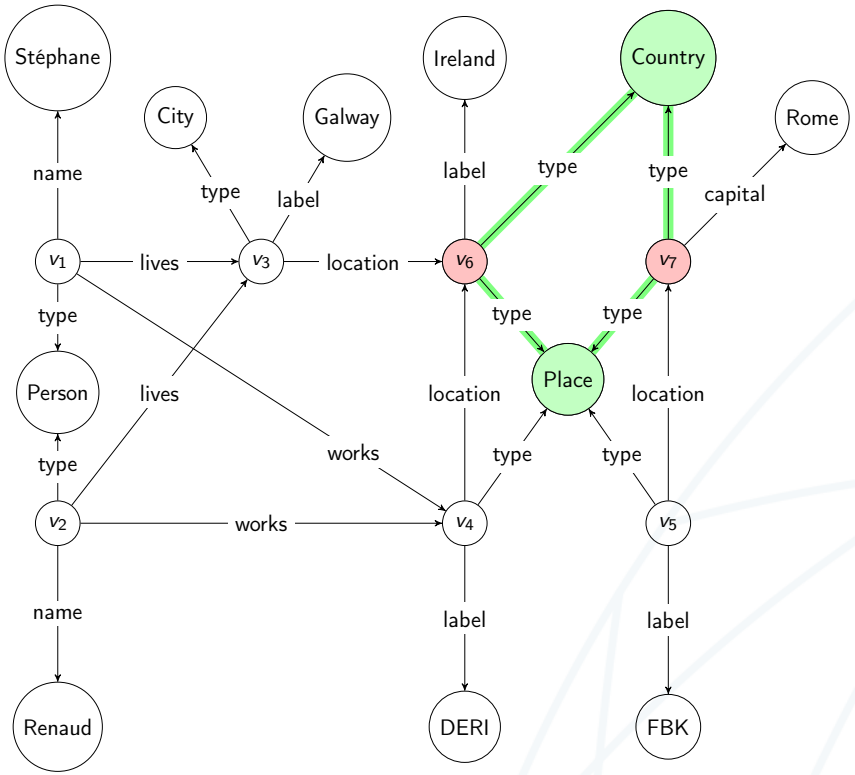


Figure: An entity graph describing people, places, and their relationships.

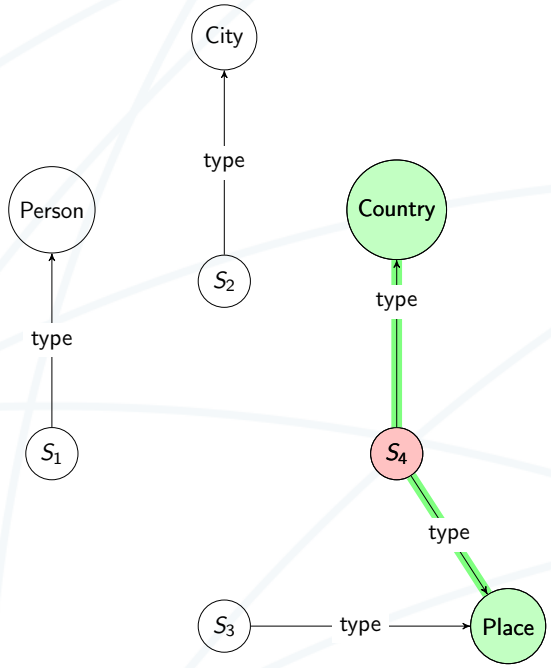


Figure: Types summary of the entity graph.

Example: Types Summary

Relation: set of types associated to a node

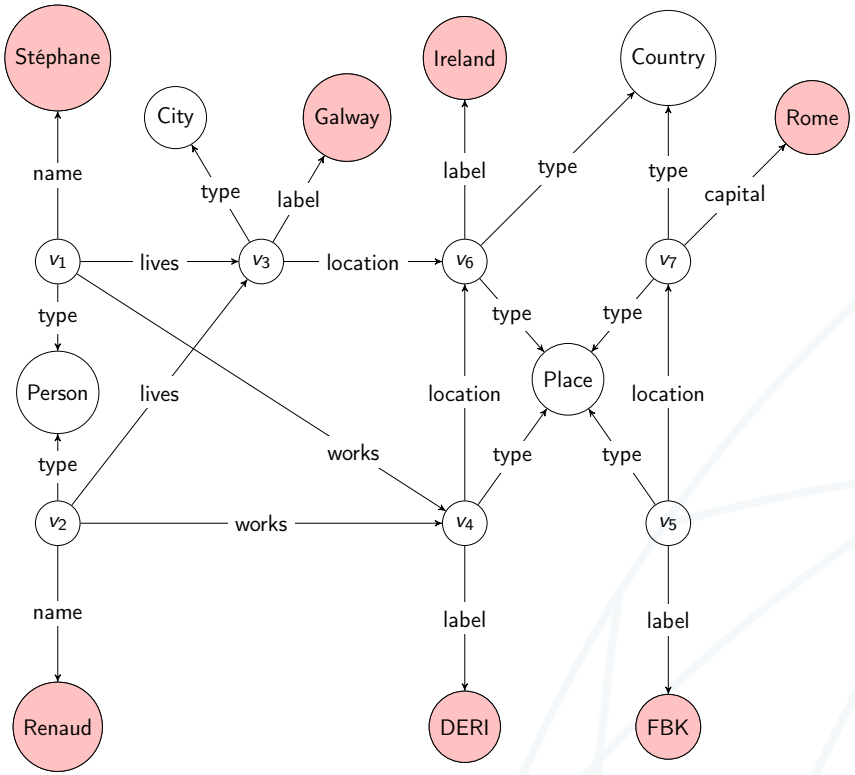


Figure: An entity graph describing people, places, and their relationships.

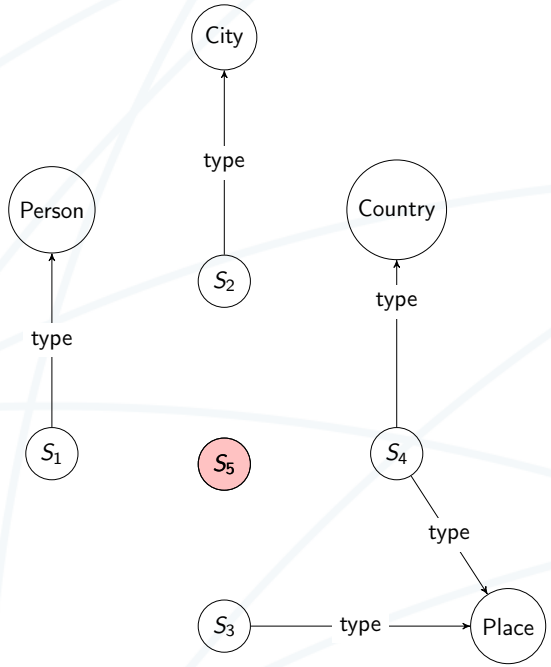


Figure: Types summary of the entity graph.

Example: Types Summary

Insight

Relation: set of types associated to a node

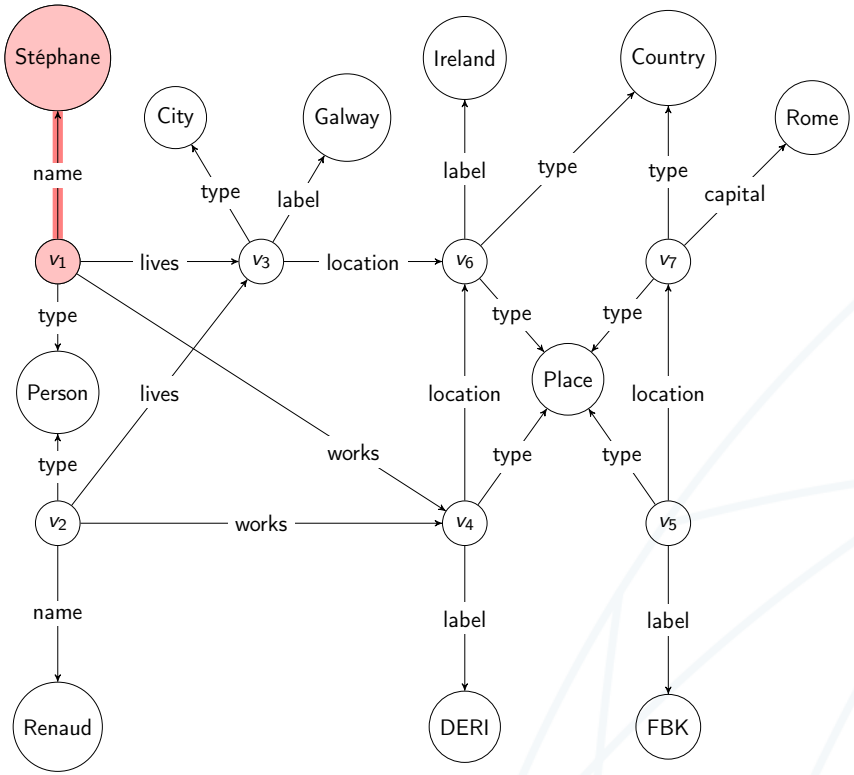


Figure: An entity graph describing people, places, and their relationships.

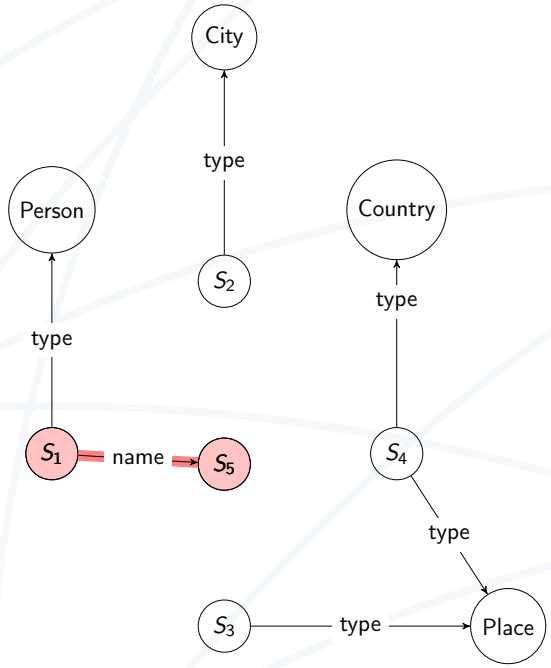


Figure: Types summary of the entity graph.

Example: Types Summary

Relation: set of types associated to a node

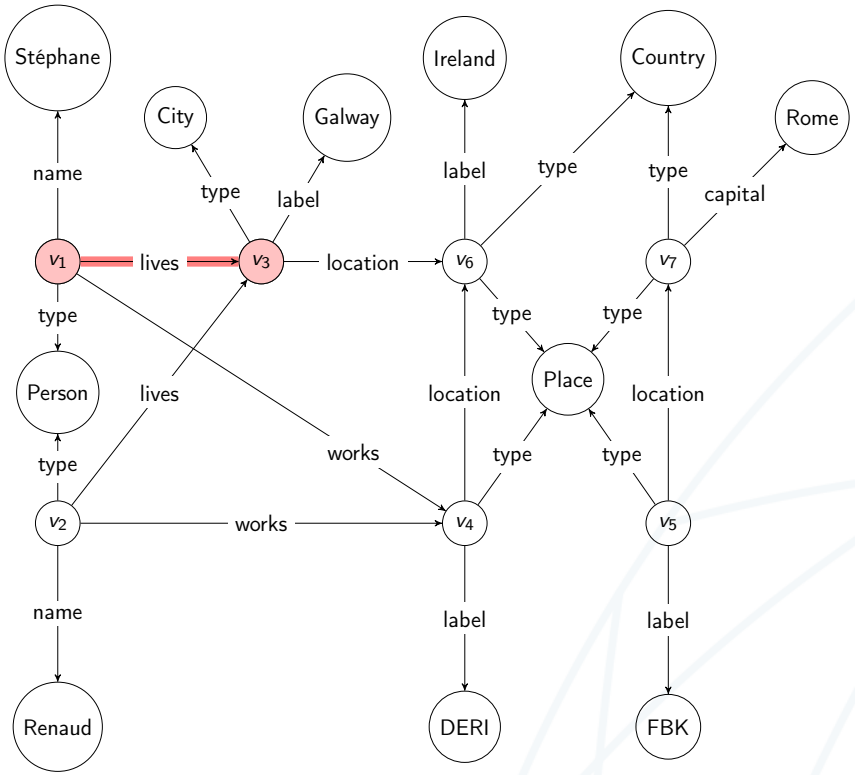


Figure: An entity graph describing people, places, and their relationships.

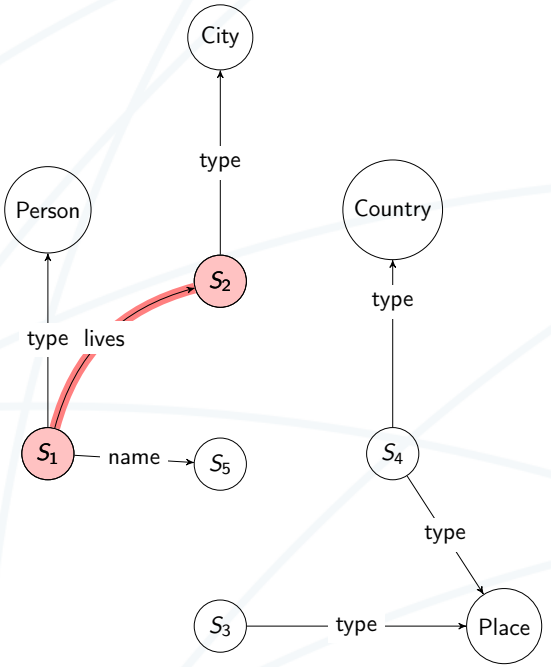


Figure: Types summary of the entity graph.

Example: Types Summary

Insight

Relation: set of types associated to a node

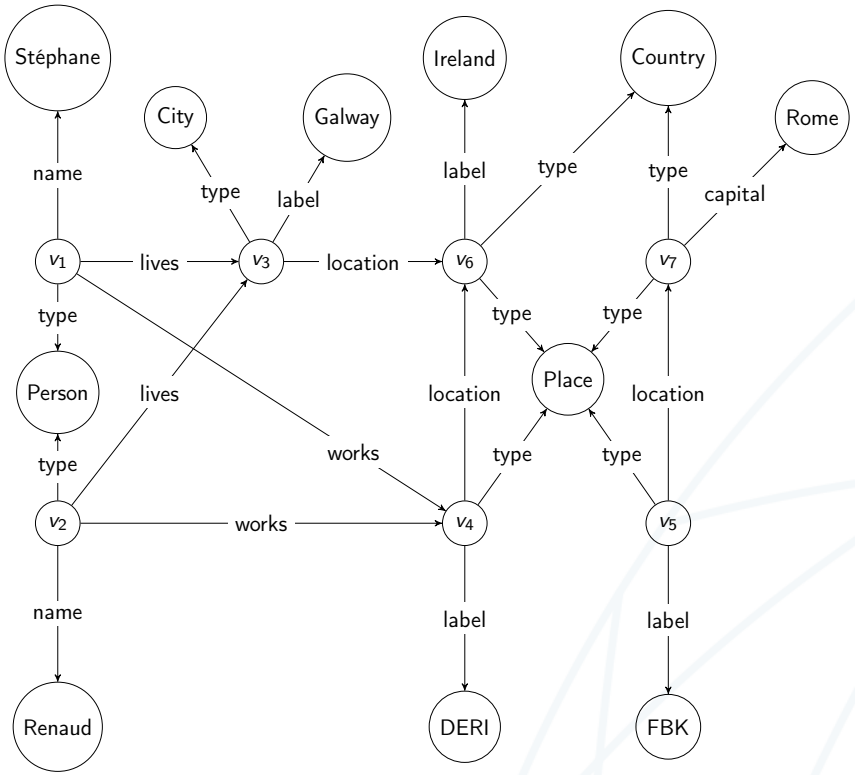


Figure: An entity graph describing people, places, and their relationships.

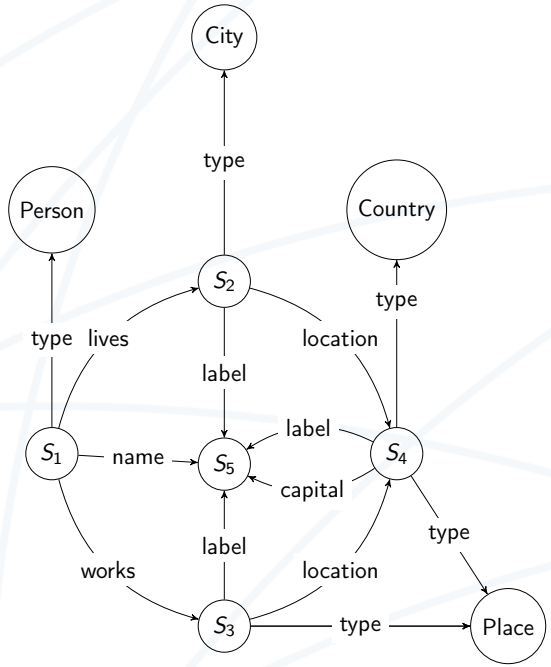


Figure: Types summary of the entity graph.

Implementations

Algorithm

1. Entity description (group operator)
2. Nodes mapping (object invention)
3. Edges materialization (join operator)

SPARQL

Performance: Timeout for graphs above 20M triples.
Pros: Leverage endpoint for optimizing queries.
Cons: Bounded by the expressivity of SPARQL.

MapReduce

Performance: Scale to large graphs up to 50B triples.
Pros: More possiblities for optimisations.
Cons: Cluster maintenance.

Implementations

Algorithm

1. Entity description (group operator)
2. Nodes mapping (object invention)
3. Edges materialization (join operator)

SPARQL

Performance: Timeout for graphs above 20M triples.
Pros: Leverage endpoint for optimizing queries.
Cons: Bounded by the expressivity of SPARQL.

MapReduce

Performance: Scale to large graphs up to 50B triples.
Pros: More possiblities for optimisations.
Cons: Cluster maintenance.

Loss of Precision with the Types Summary

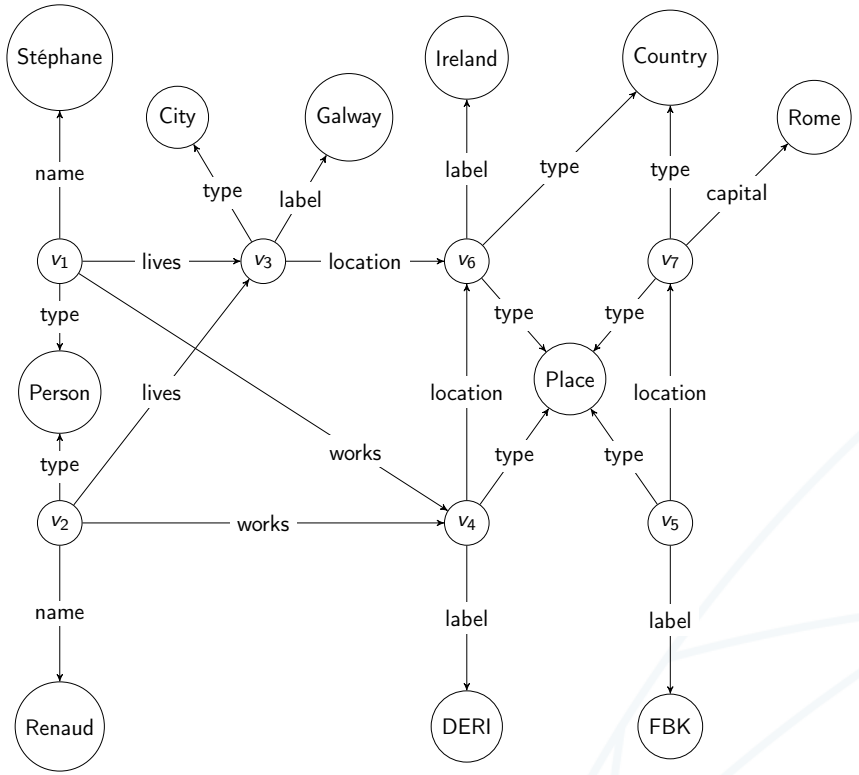


Figure: An entity graph describing people, places, and their relationships.

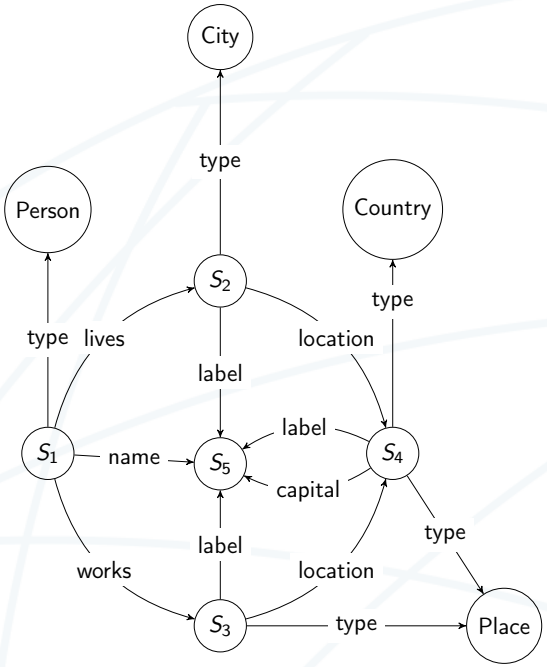


Figure: Types summary of the entity graph.

Loss of Precision with the Types Summary

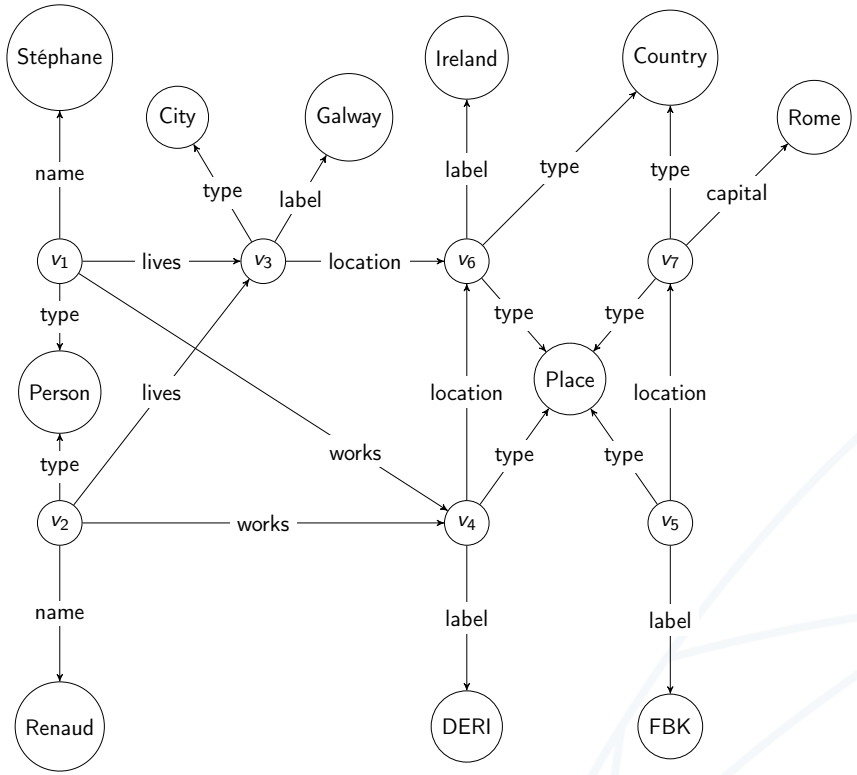


Figure: An entity graph describing people, places, and their relationships.

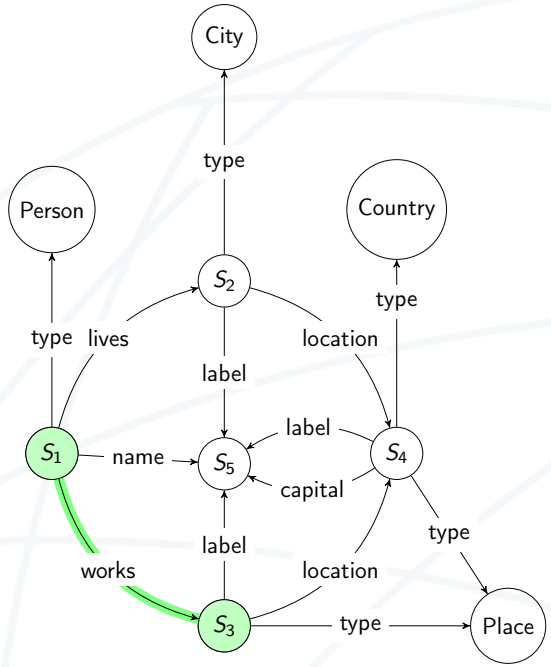


Figure: Types summary of the entity graph.

Loss of Precision with the Types Summary

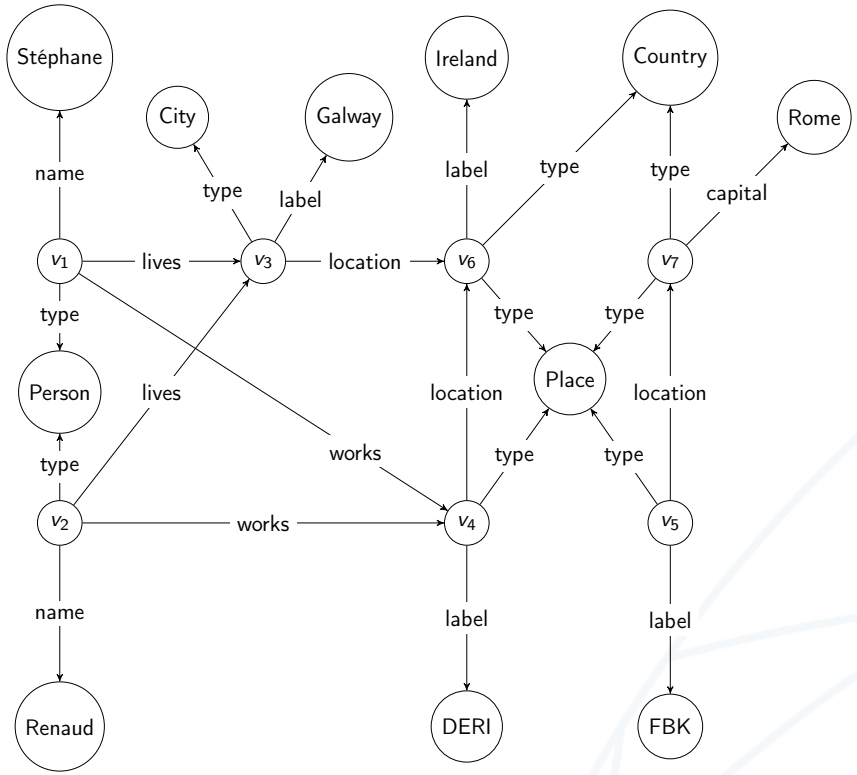


Figure: An entity graph describing people, places, and their relationships.

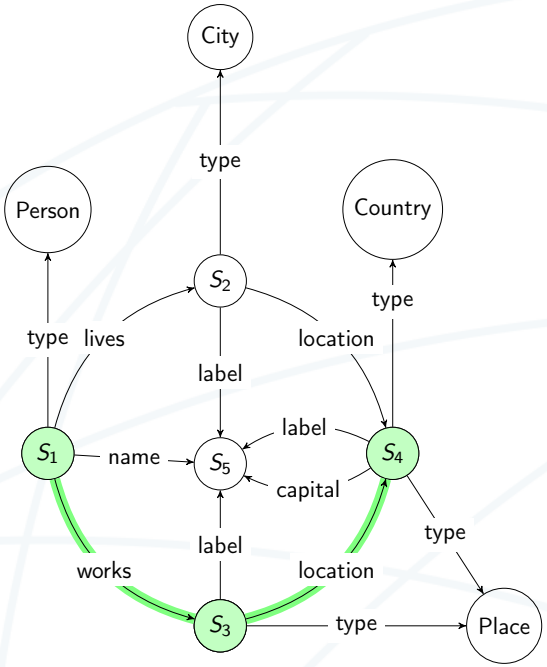


Figure: Types summary of the entity graph.

Loss of Precision with the Types Summary

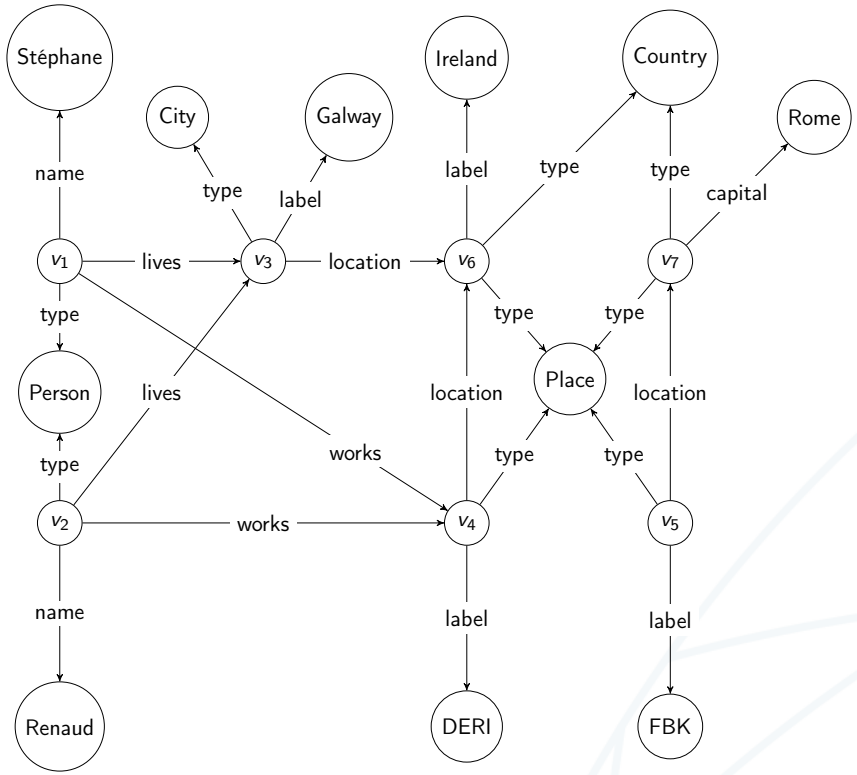


Figure: An entity graph describing people, places, and their relationships.

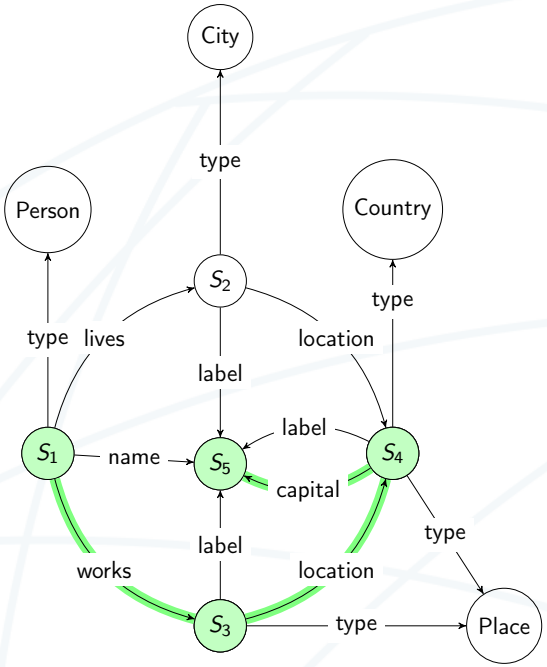


Figure: Types summary of the entity graph.

Loss of Precision with the Types Summary

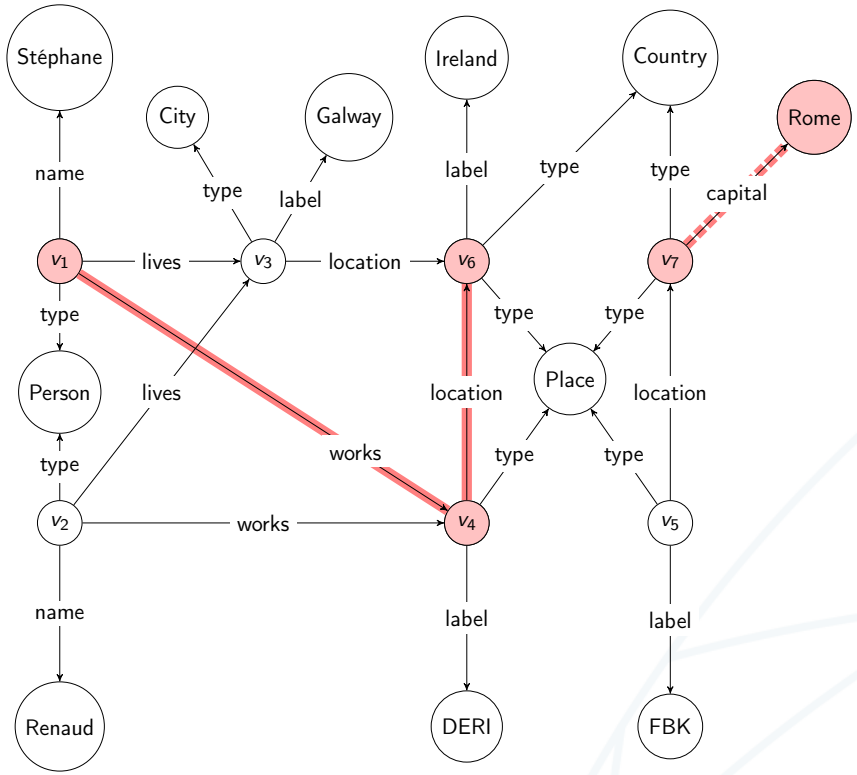


Figure: An entity graph describing people, places, and their relationships.

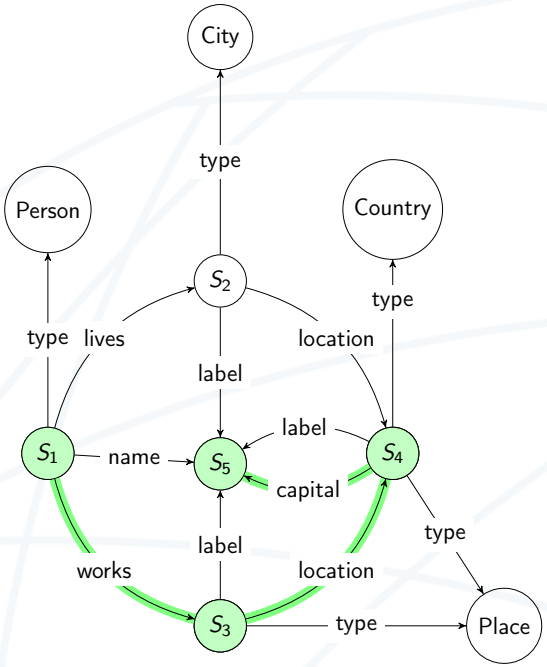


Figure: Types summary of the entity graph.

Loss of Precision with the Types Summary

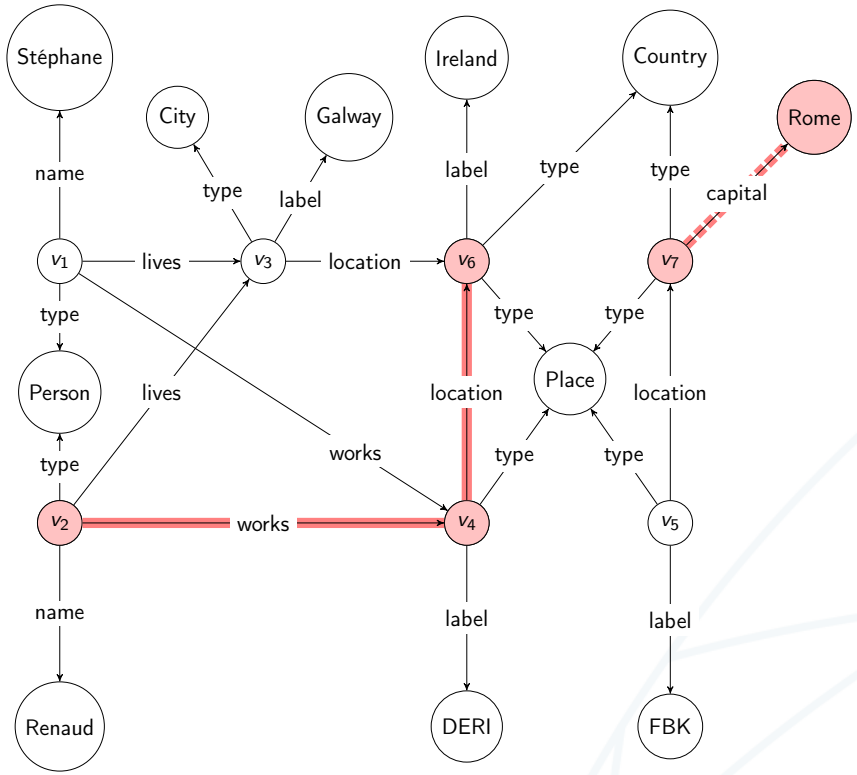


Figure: An entity graph describing people, places, and their relationships.

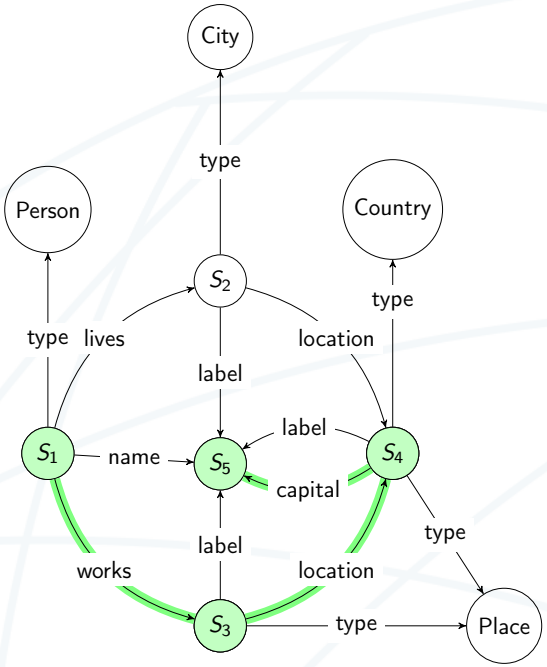
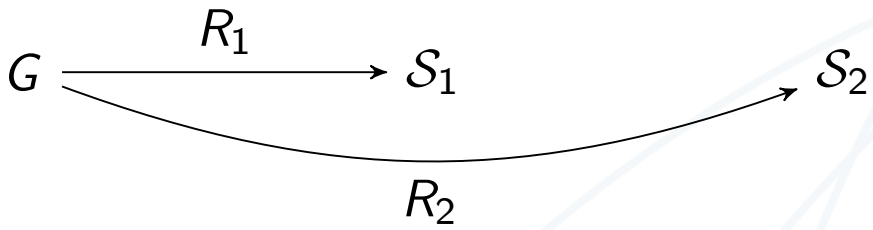


Figure: Types summary of the entity graph.

Ordering of Graph Summaries

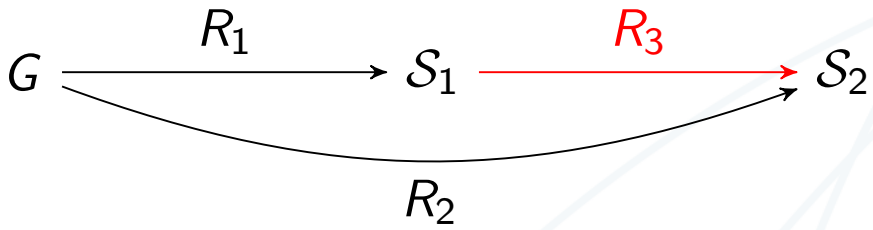
Partial Relation



Graph Lattice

Ordering of Graph Summaries

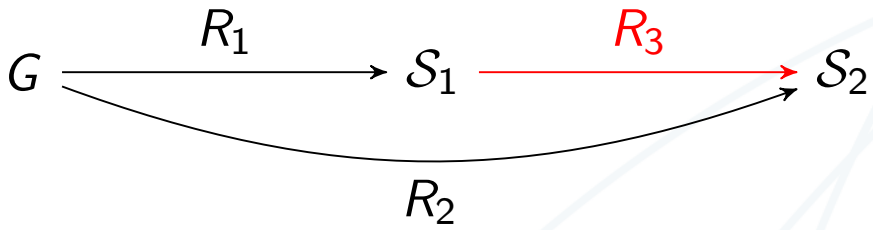
Partial Relation



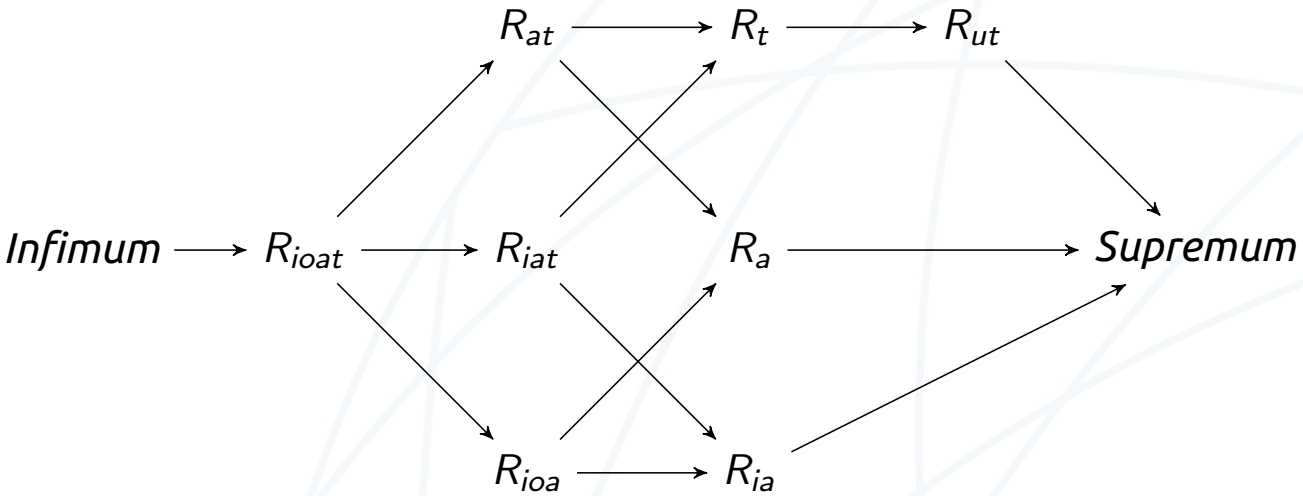
Graph Lattice

Ordering of Graph Summaries

Partial Relation



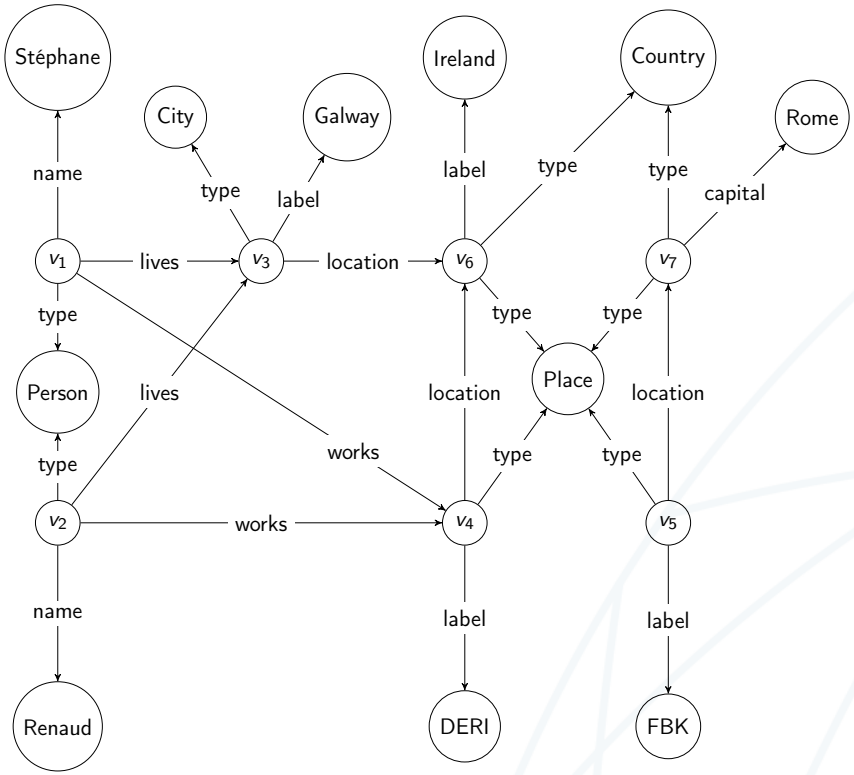
Graph Lattice

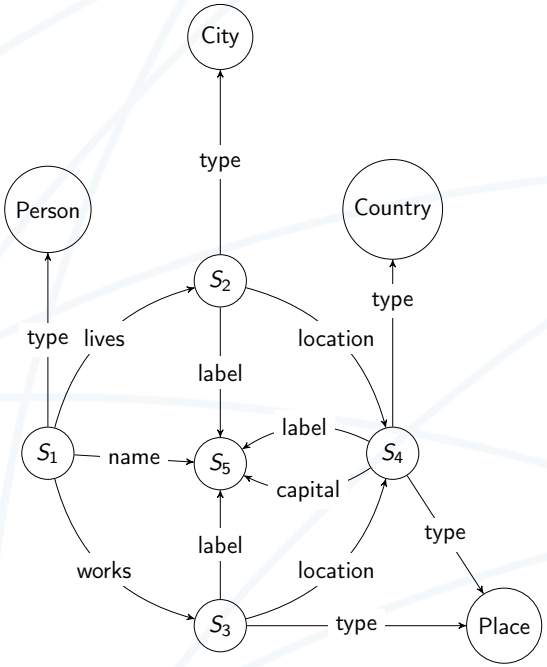


Comparison

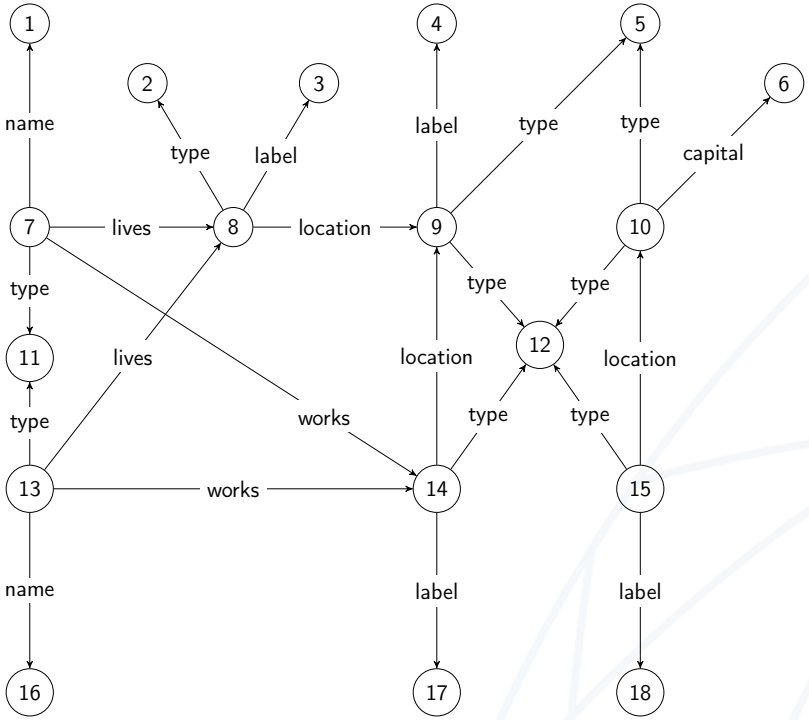
- Leverage the partial relation for ordering the summaries
- Compare a summary against the original graph
- Compute how many errors the summarisation committed

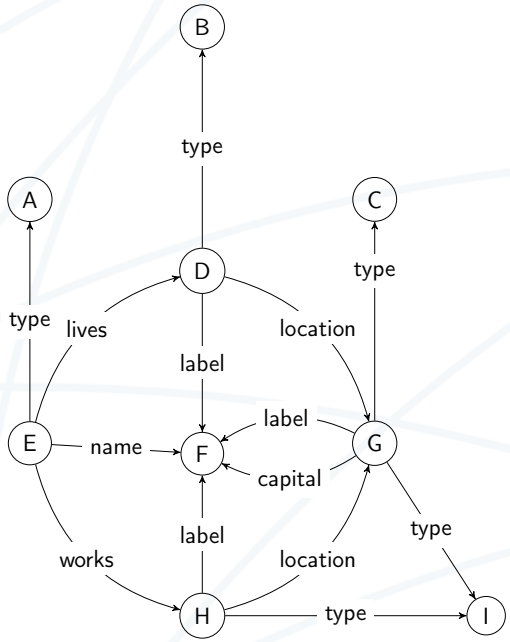
An Example:
Evaluation of the Summary Precision





An Example:
Evaluation of the Summary Precision

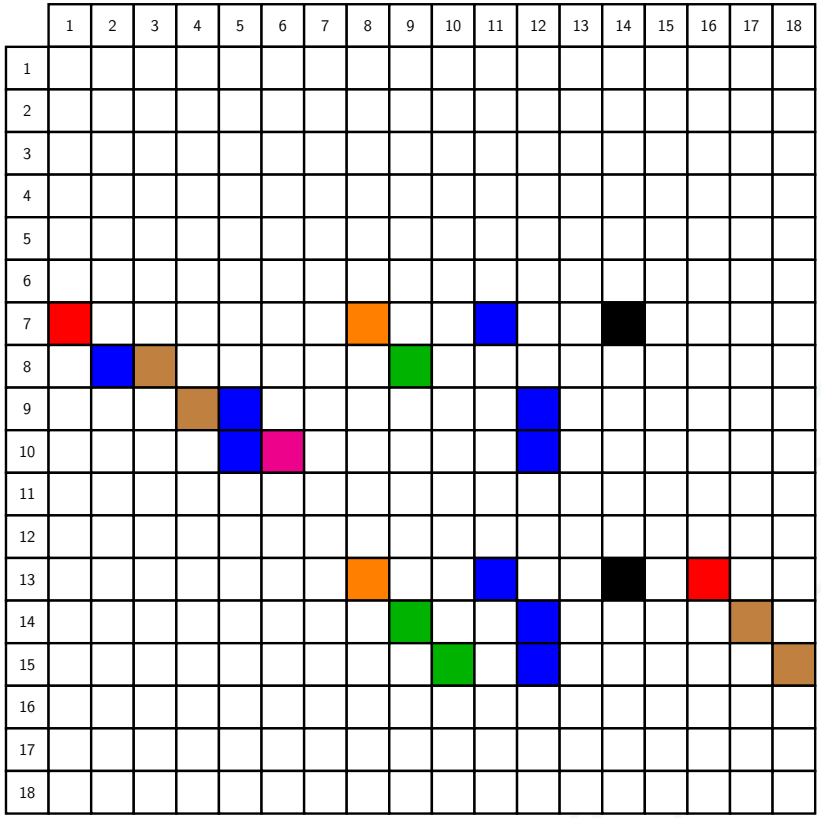


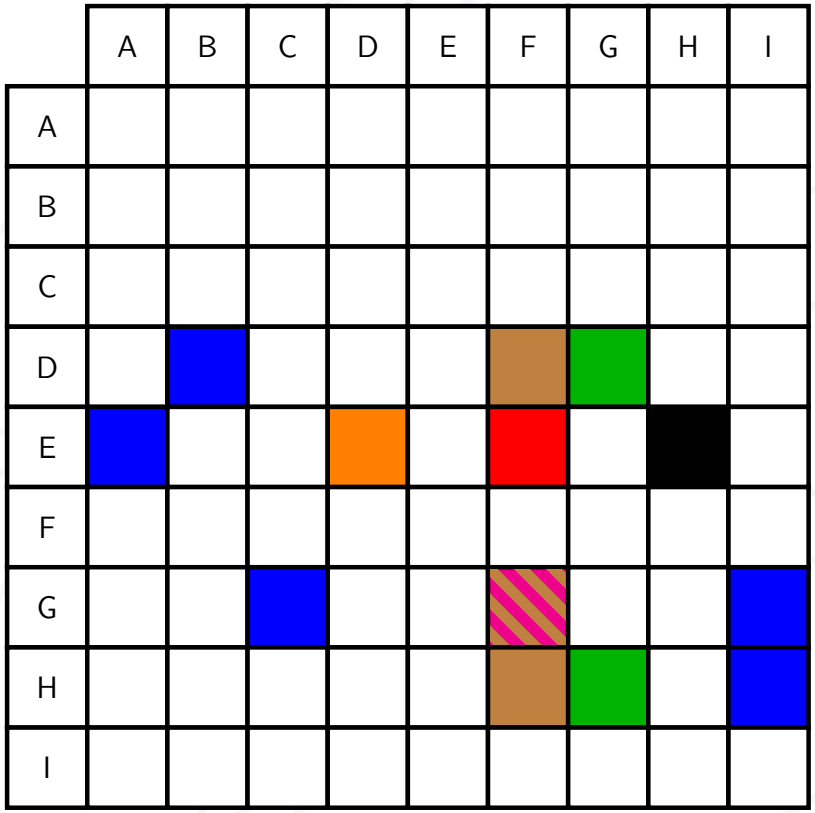


An Example:
Evaluation of the Summary Precision

Insight

name type label capital lives location works

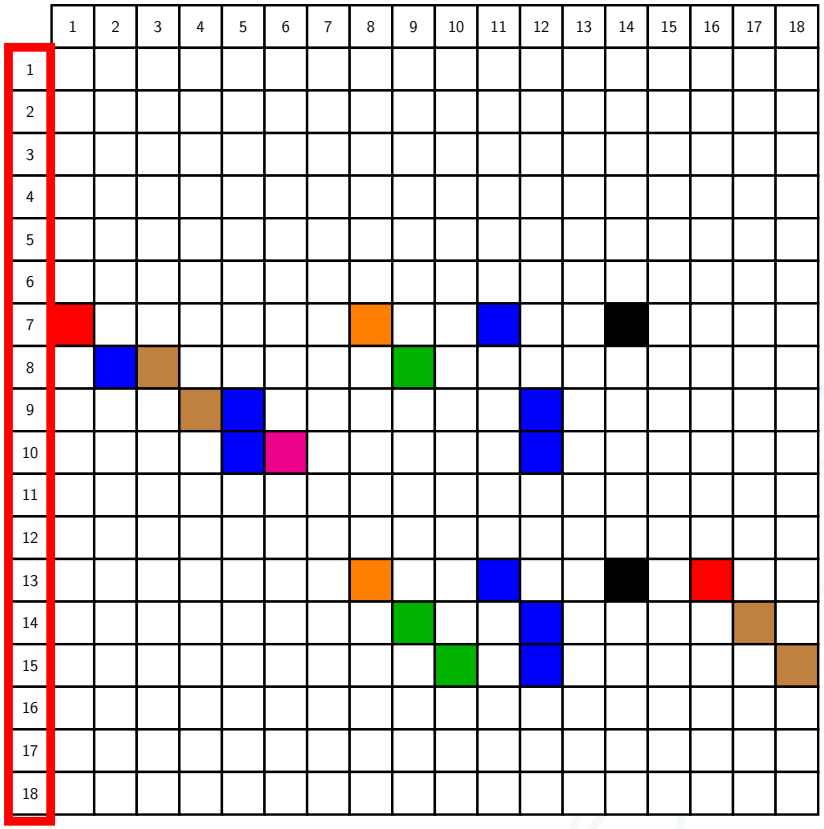


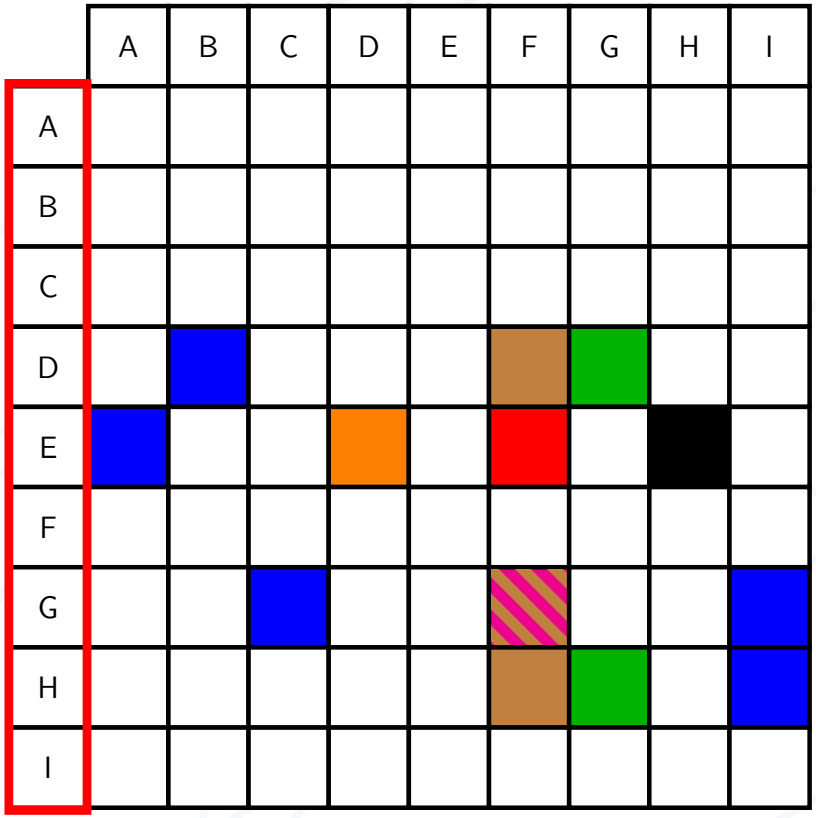


An Example:
Evaluation of the Summary Precision

Insight

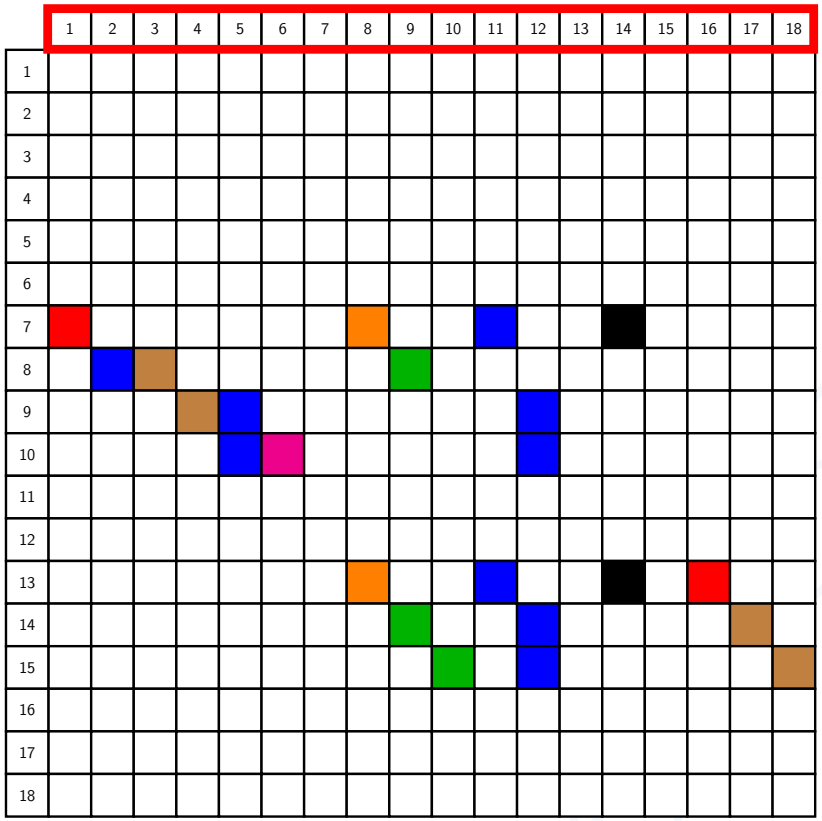
name type label capital lives location works

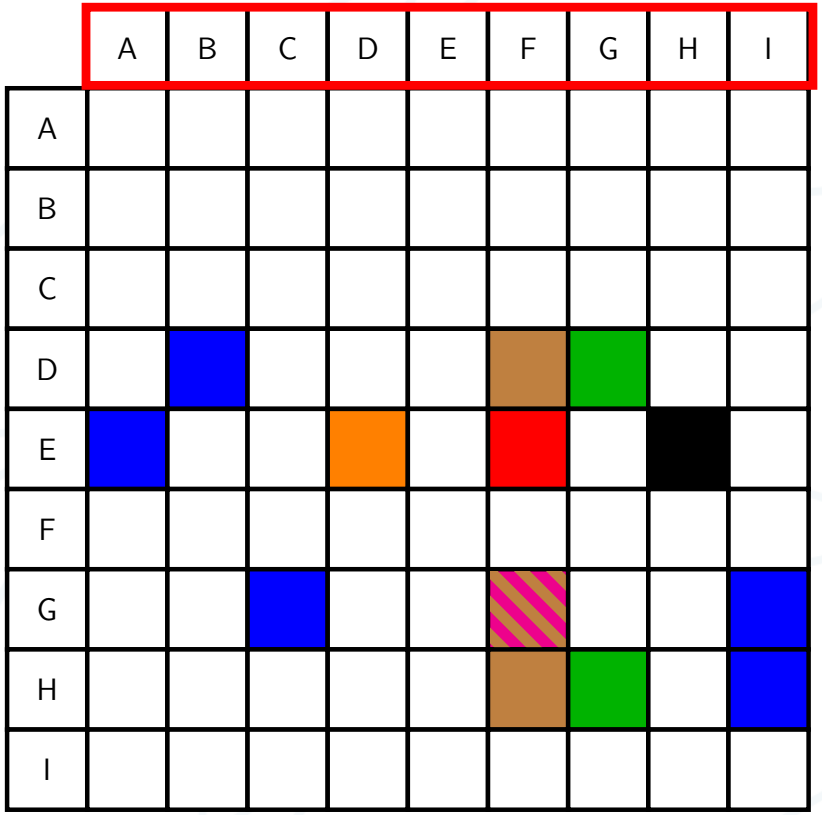




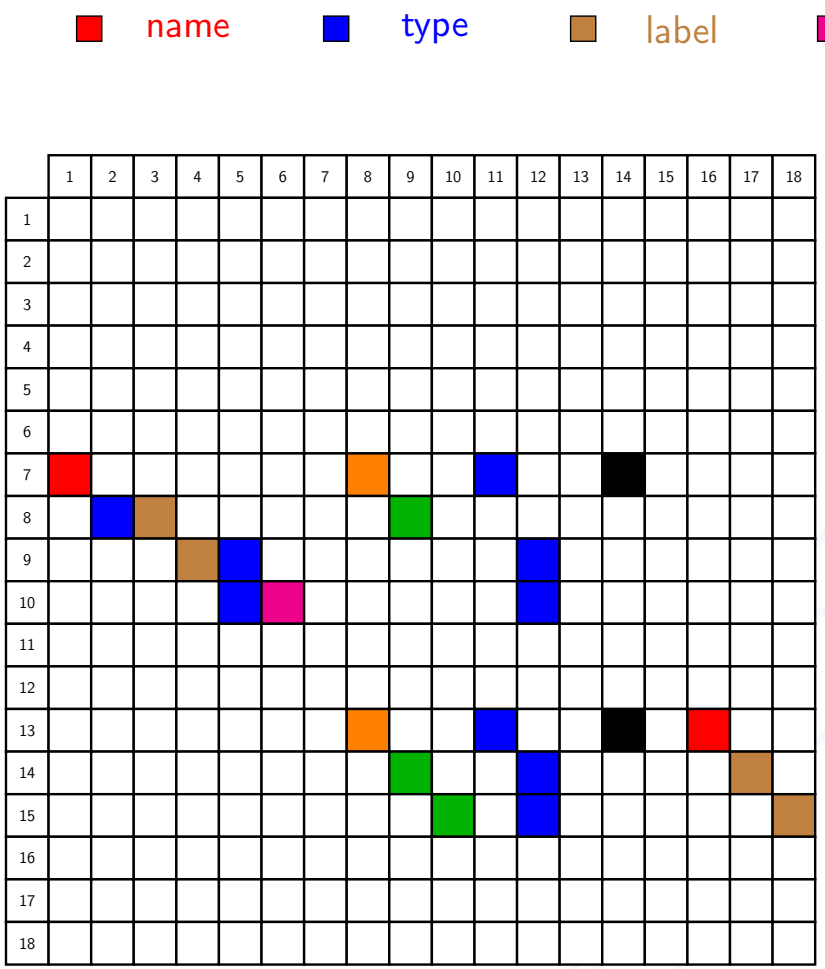
An Example:
Evaluation of the Summary Precision

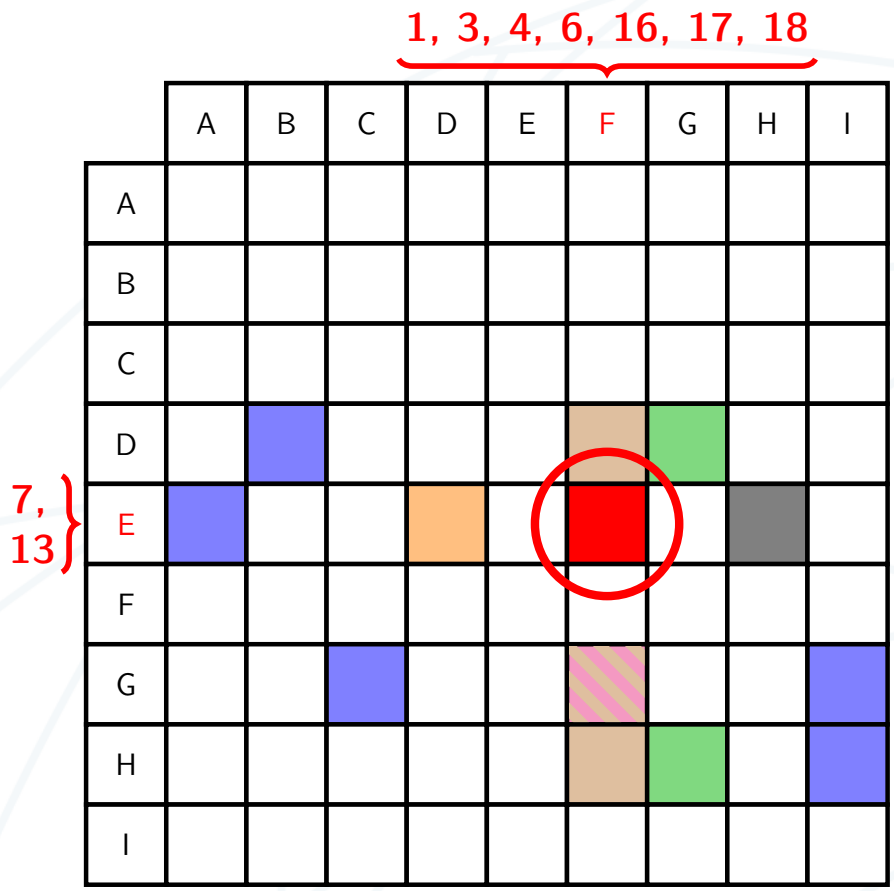
name type label capital lives location works



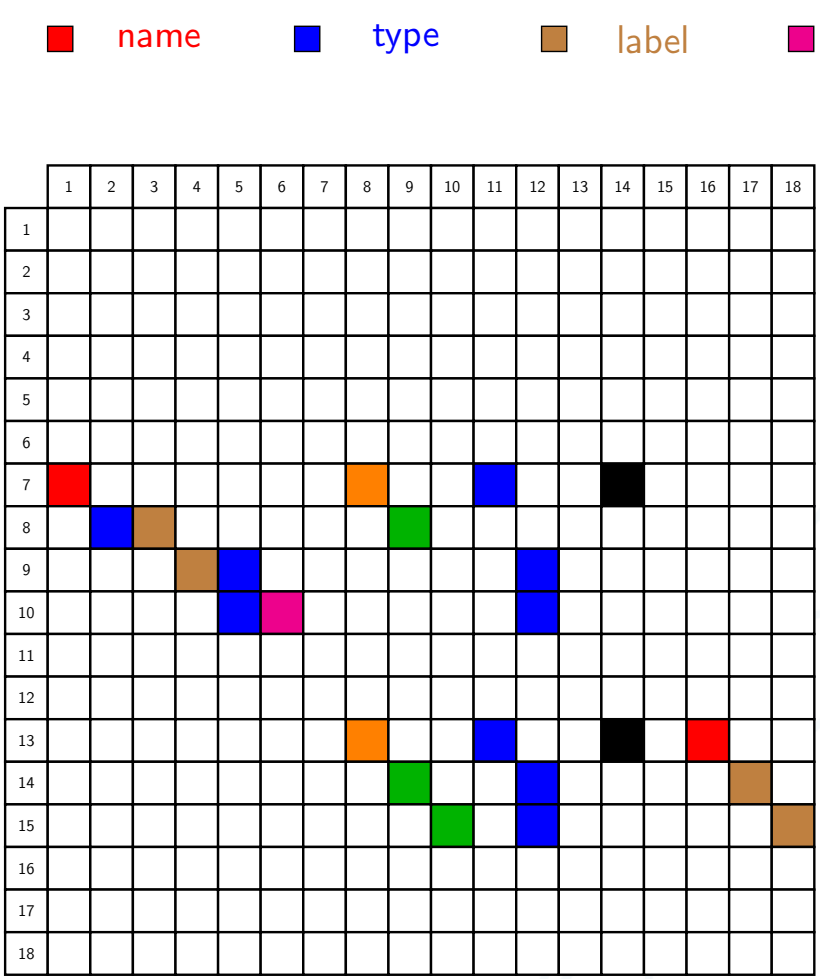


An Example:
Evaluation of the Summary Precision

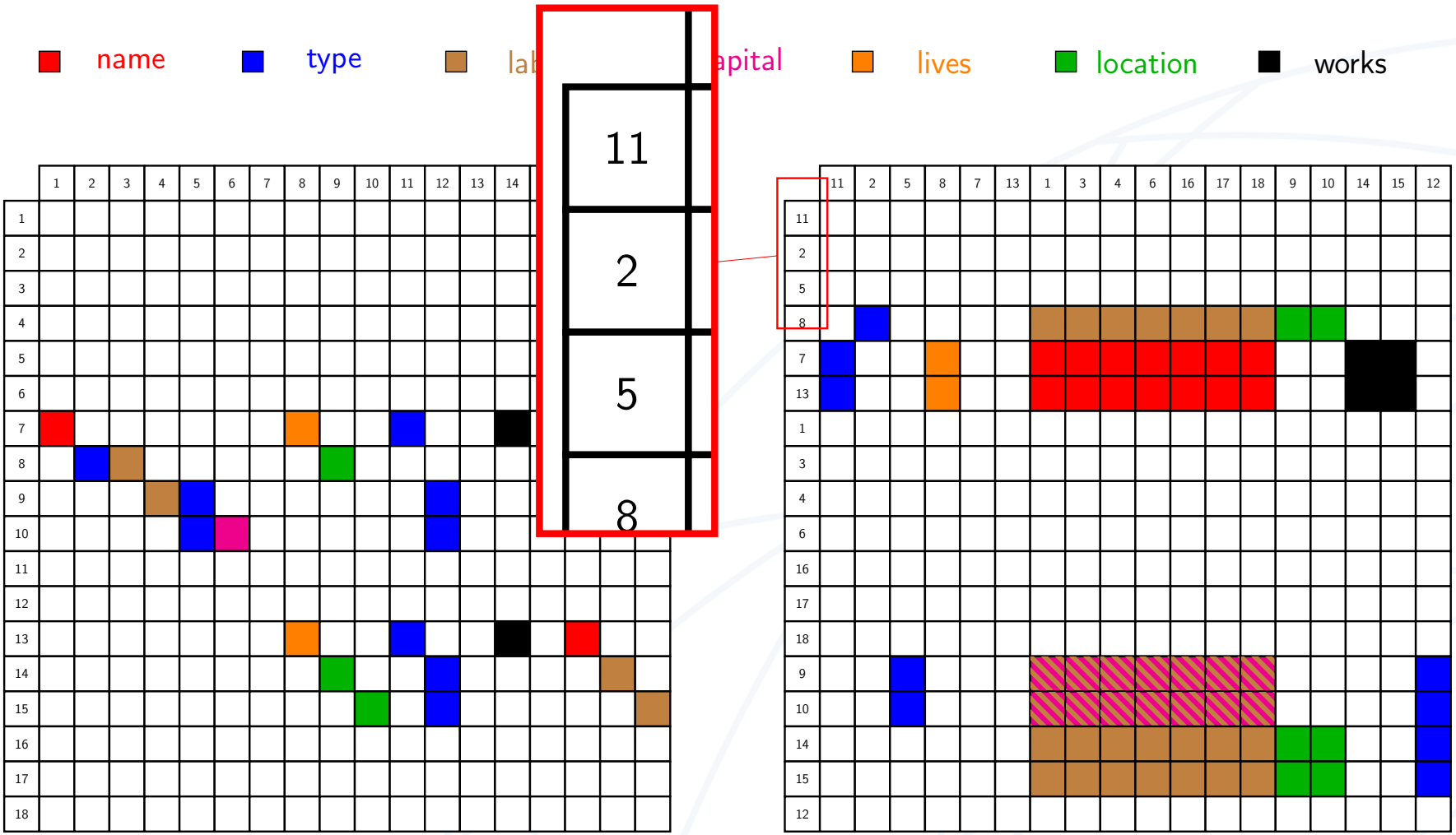




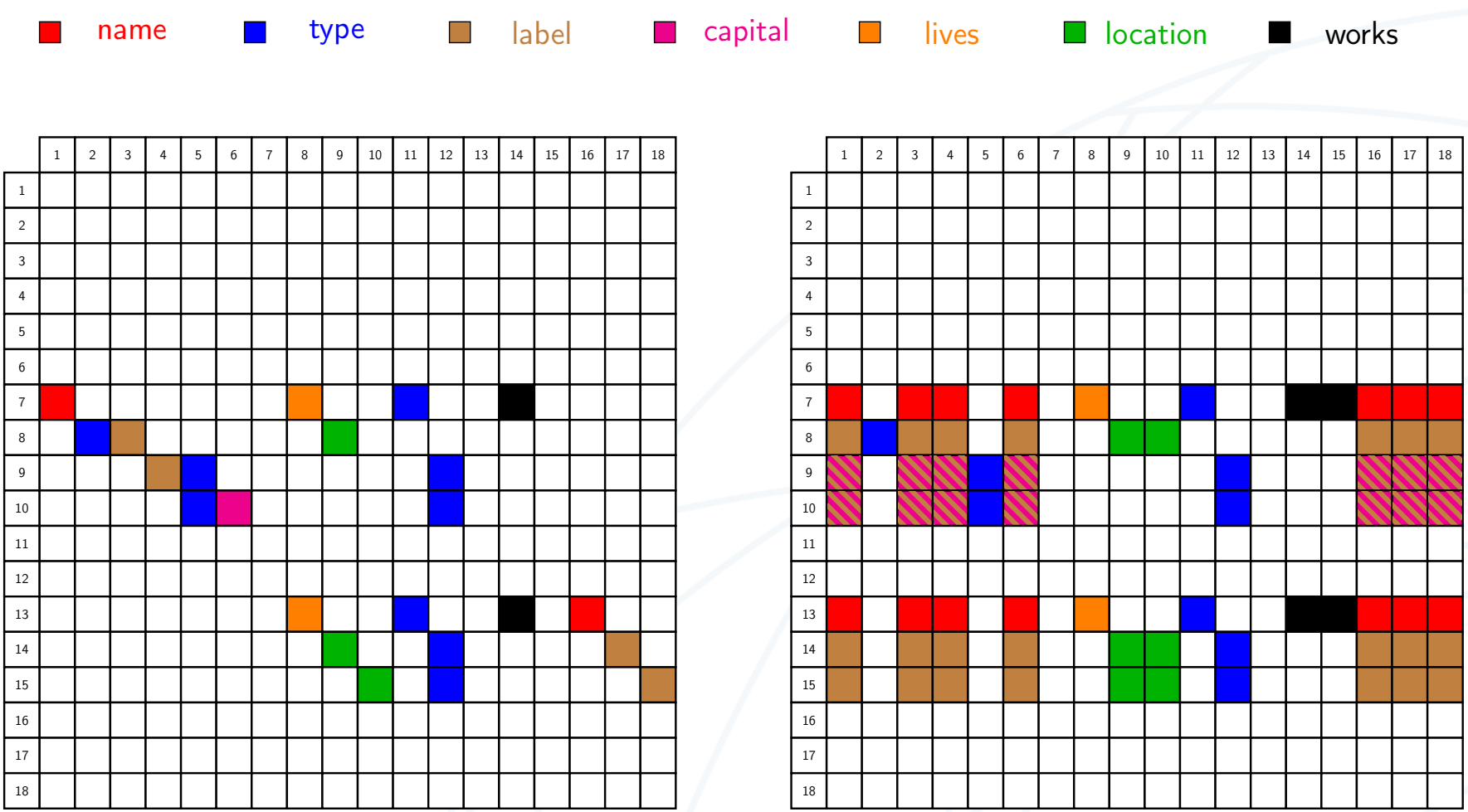
An Example:
Evaluation of the Summary Precision



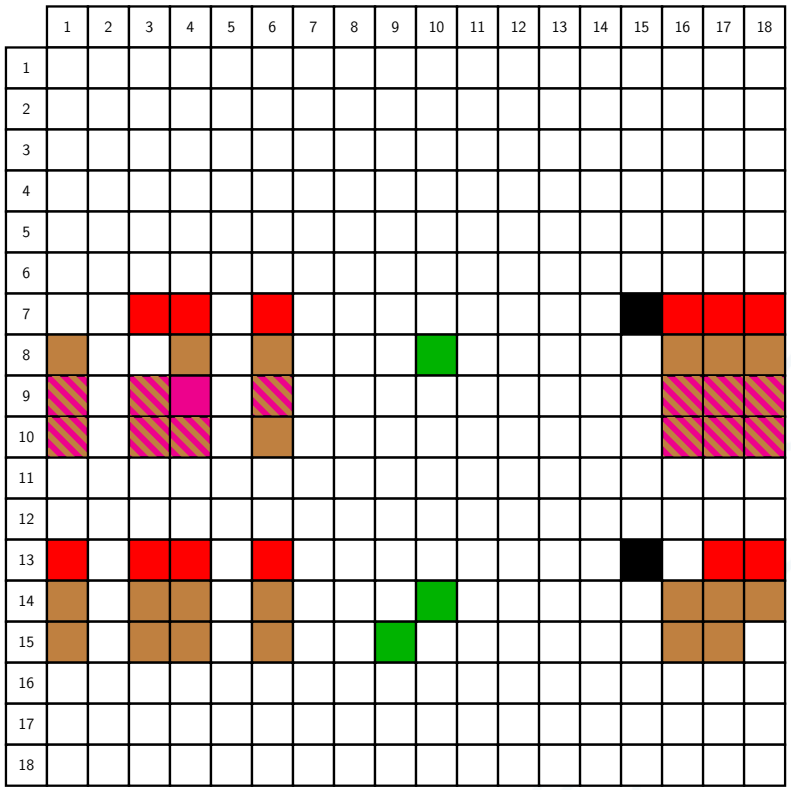
An Example:
Evaluation of the Summary Precision

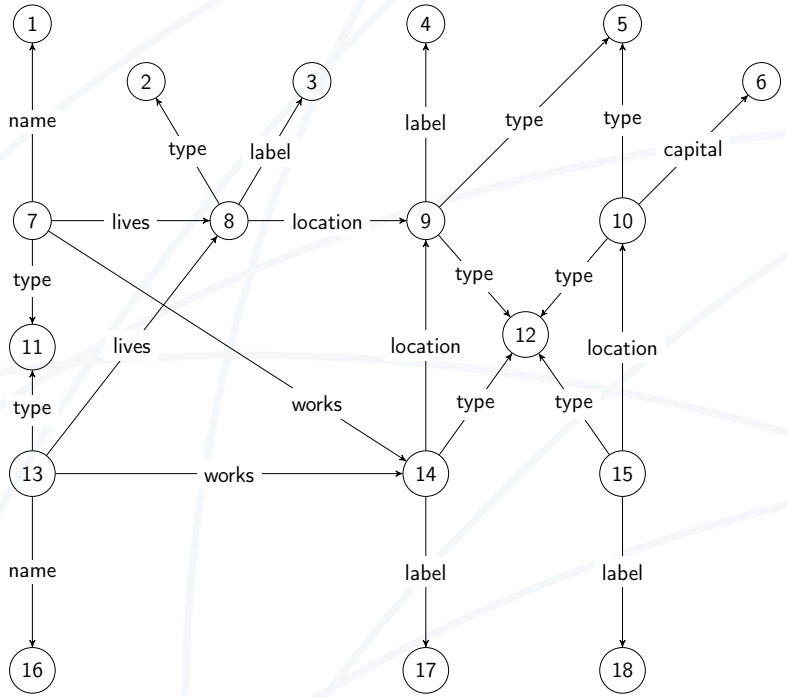


An Example:
Evaluation of the Summary Precision

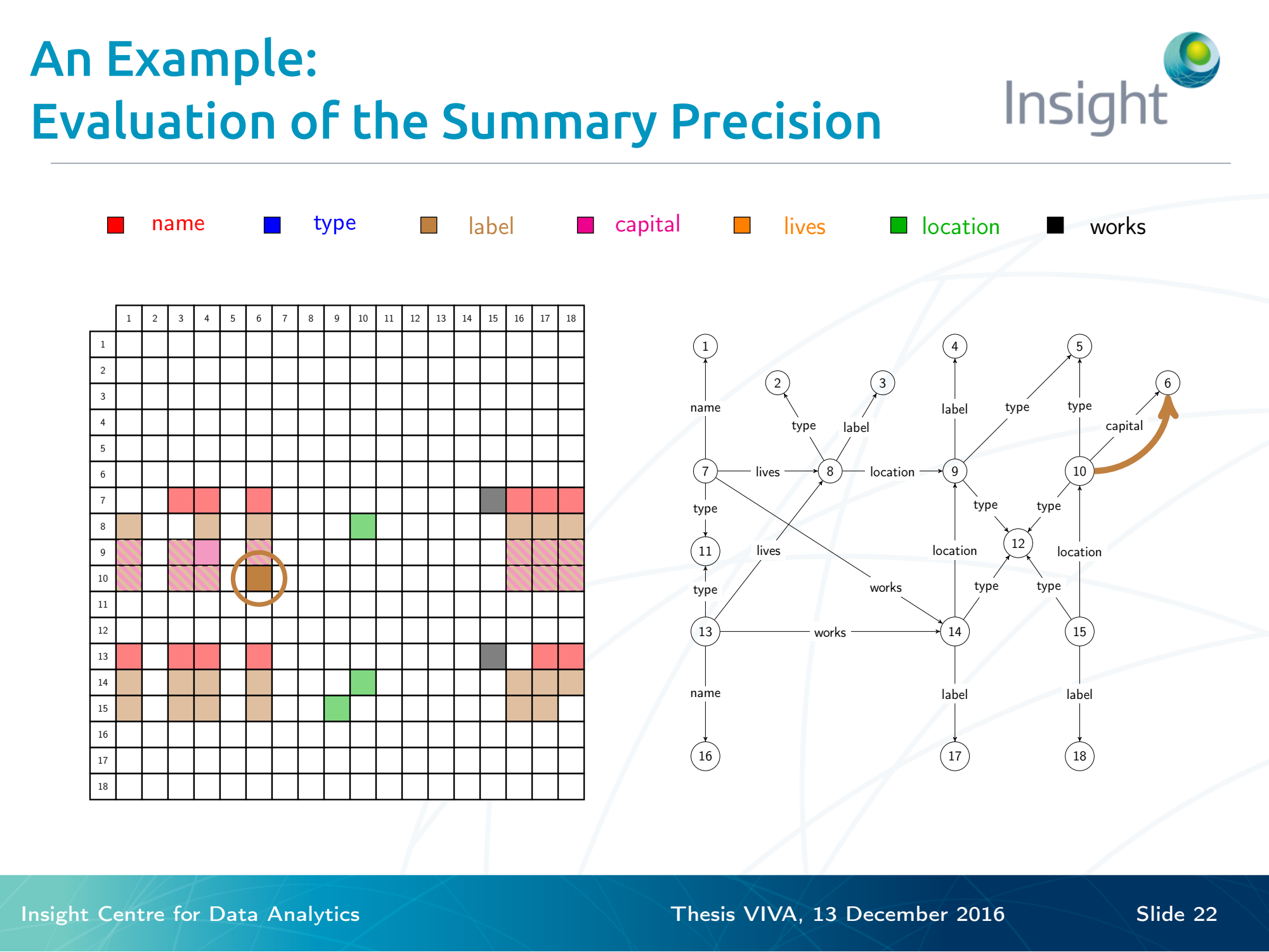


An Example: Evaluation of the Summary Precision



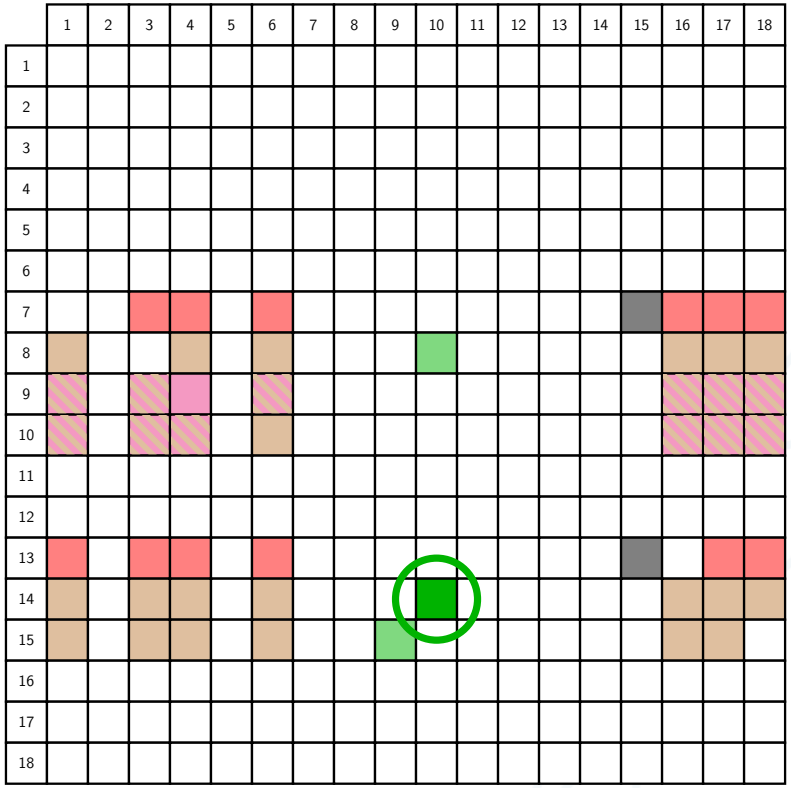


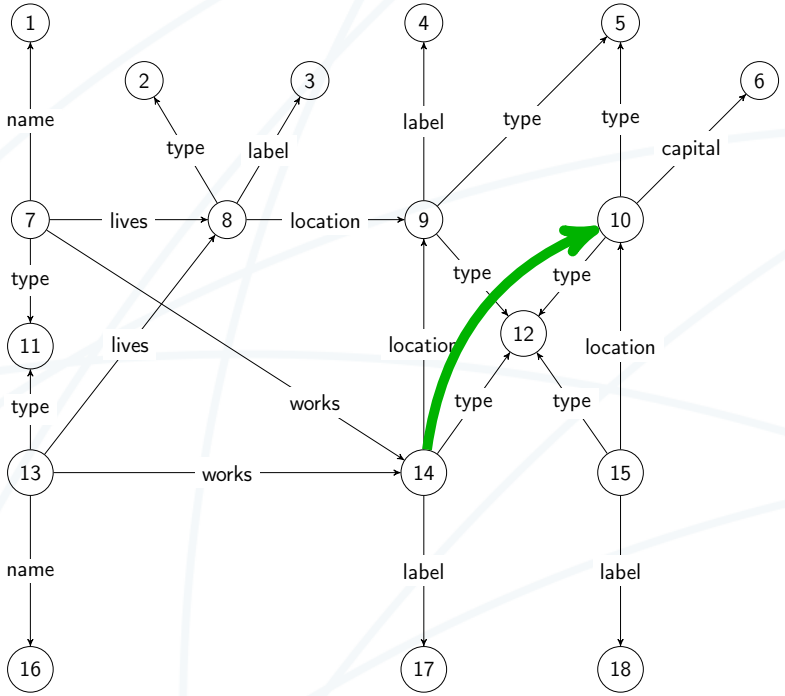
An Example:
Evaluation of the Summary Precision



An Example:
Evaluation of the Summary Precision

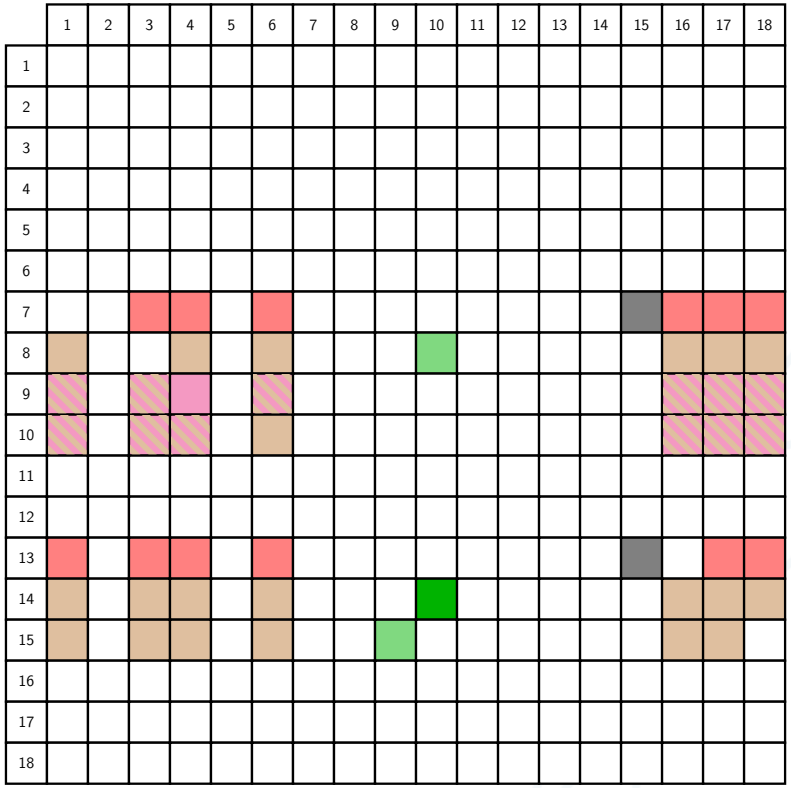
name type label capital lives location works

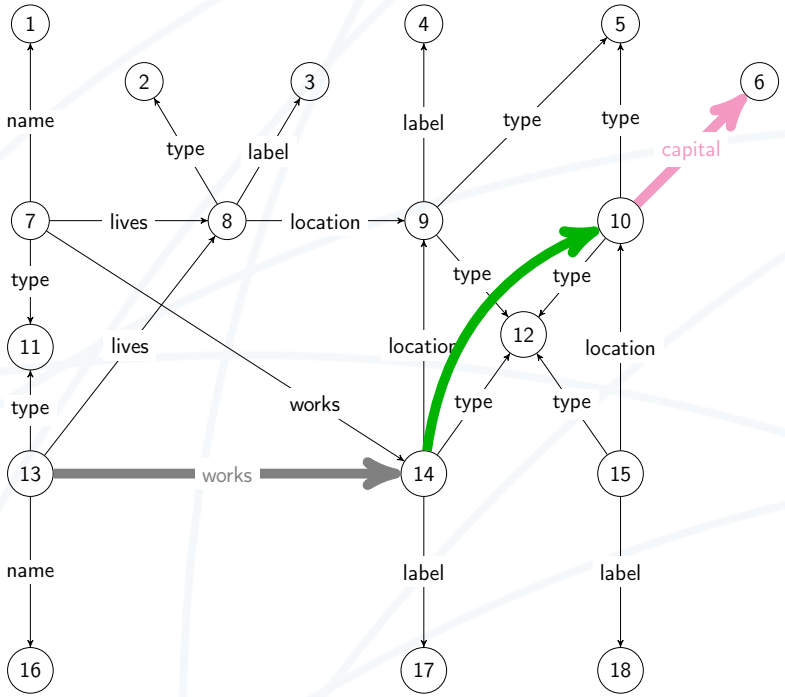




An Example:
Evaluation of the Summary Precision

name type label capital lives location works





Summary Precision

Errors Classification

- Type
- Attribute
- Connectivity

Precision Measure

- Based on the set of **true** (TP) and **false** (FP) positives edges:

$$Prec(R, x) = \frac{|TP(x)|}{|TP(x) \bigcup FP(R, x)|}$$

Evaluation Dimensions

How small the summary is compared to the original graph ?

Ratio of order and size of the graph G to the summary $\mathcal{S}$:

$$G : \mathcal{S} = \frac{|V| + |A|}{|\mathcal{W}| + |\mathcal{B}|}$$

How performant and scalable is the summarization process ?

Record of the time spent on the process

How precise is the generated summary ?

Average precision of the evaluated summary based on a gold-standard summary:

$$\frac{1}{|\mathcal{W}_{fbt}|} \times \sum_{c \in \mathcal{W}} \frac{1}{|C(c)|} \times \sum_{x \in C(c)} Prec(R, x)$$

Evaluation Dimensions

How small the summary is compared to the original graph ?

Ratio of order and size of the graph G to the summary $\mathcal{S}$:

$$G : \mathcal{S} = \frac{|V| + |A|}{|\mathcal{W}| + |\mathcal{B}|}$$

How performant and scalable is the summarization process ?

Record of the time spent on the process

How precise is the generated summary ?

Average precision of the evaluated summary based on a gold-standard summary:

$$\frac{1}{|\mathcal{W}_{fbt}|} \times \sum_{c \in \mathcal{W}} \frac{1}{|C(c)|} \times \sum_{x \in C(c)} Prec(R, x)$$

Evaluation Dimensions

How small the summary is compared to the original graph ?

Ratio of order and size of the graph G to the summary $\mathcal{S}$:

$$G : \mathcal{S} = \frac{|V| + |A|}{|\mathcal{W}| + |\mathcal{B}|}$$

How performant and scalable is the summarization process ?

Record of the time spent on the process

How precise is the generated summary ?

Average precision of the evaluated summary based on a gold-standard summary:

$$\frac{1}{|\mathcal{W}_{fbt}|} \times \sum_{c \in \mathcal{W}} \frac{1}{|C(c)|} \times \sum_{x \in C(c)} Prec(R, x)$$

Evaluation Dimensions

How small the summary is compared to the original graph ?

Ratio of order and size of the graph G to the summary $\mathcal{S}$:

$$G : \mathcal{S} = \frac{|V| + |A|}{|\mathcal{W}| + |\mathcal{B}|}$$

How performant and scalable is the summarization process ?

Record of the time spent on the process

How precise is the generated summary ?

Average precision of the evaluated summary based on a gold-standard summary:

$$\frac{1}{|\mathcal{W}_{fbt}|} \times \sum_{c \in \mathcal{W}} \frac{1}{|C(c)|} \times \sum_{x \in C(c)} Prec(R, x)$$

Evaluation Dimensions

How small the summary is compared to the original graph ?

Ratio of order and size of the graph G to the summary $\mathcal{S}$:

$$G : \mathcal{S} = \frac{|V| + |A|}{|\mathcal{W}| + |\mathcal{B}|}$$

How performant and scalable is the summarization process ?

Record of the time spent on the process

How precise is the generated summary ?

Average precision of the evaluated summary based on a gold-standard summary:

$$\frac{1}{|\mathcal{W}_{fbt}|} \times \sum_{c \in \mathcal{W}} \frac{1}{|C(c)|} \times \sum_{x \in C(c)} Prec(R, x)$$

Evaluation Dimensions

How small the summary is compared to the original graph ?

Ratio of order and size of the graph G to the summary $\mathcal{S}$:

$$G : \mathcal{S} = \frac{|V| + |A|}{|\mathcal{W}| + |\mathcal{B}|}$$

How performant and scalable is the summarization process ?

Record of the time spent on the process

How precise is the generated summary ?

Average precision of the evaluated summary based on a gold-standard summary:

$$\frac{1}{|\mathcal{W}_{fbt}|} \times \sum_{c \in \mathcal{W}} \boxed{\frac{1}{|C(c)|} \times \sum_{x \in C(c)} Prec(R, x)}$$

Evaluation Dimensions

How small the summary is compared to the original graph ?

Ratio of order and size of the graph G to the summary $\mathcal{S}$:

$$G : \mathcal{S} = \frac{|V| + |A|}{|\mathcal{W}| + |\mathcal{B}|}$$

How performant and scalable is the summarization process ?

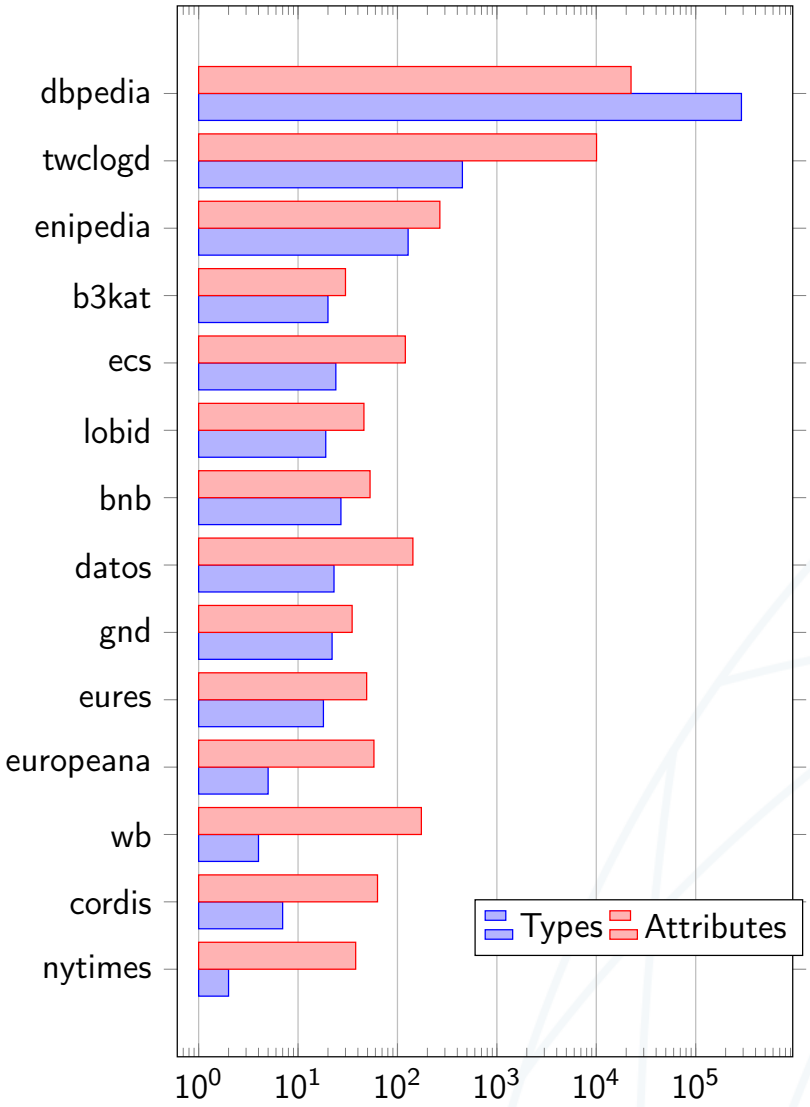
Record of the time spent on the process

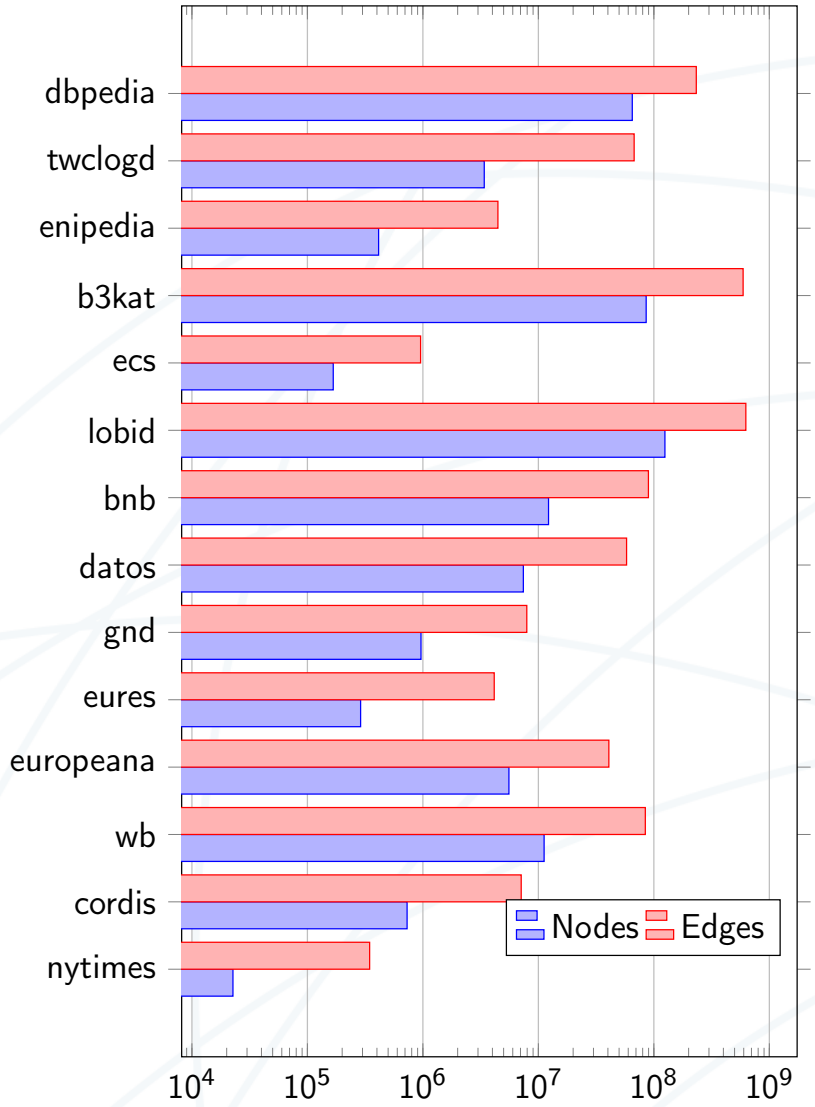
How precise is the generated summary ?

Average precision of the evaluated summary based on a gold-standard summary:

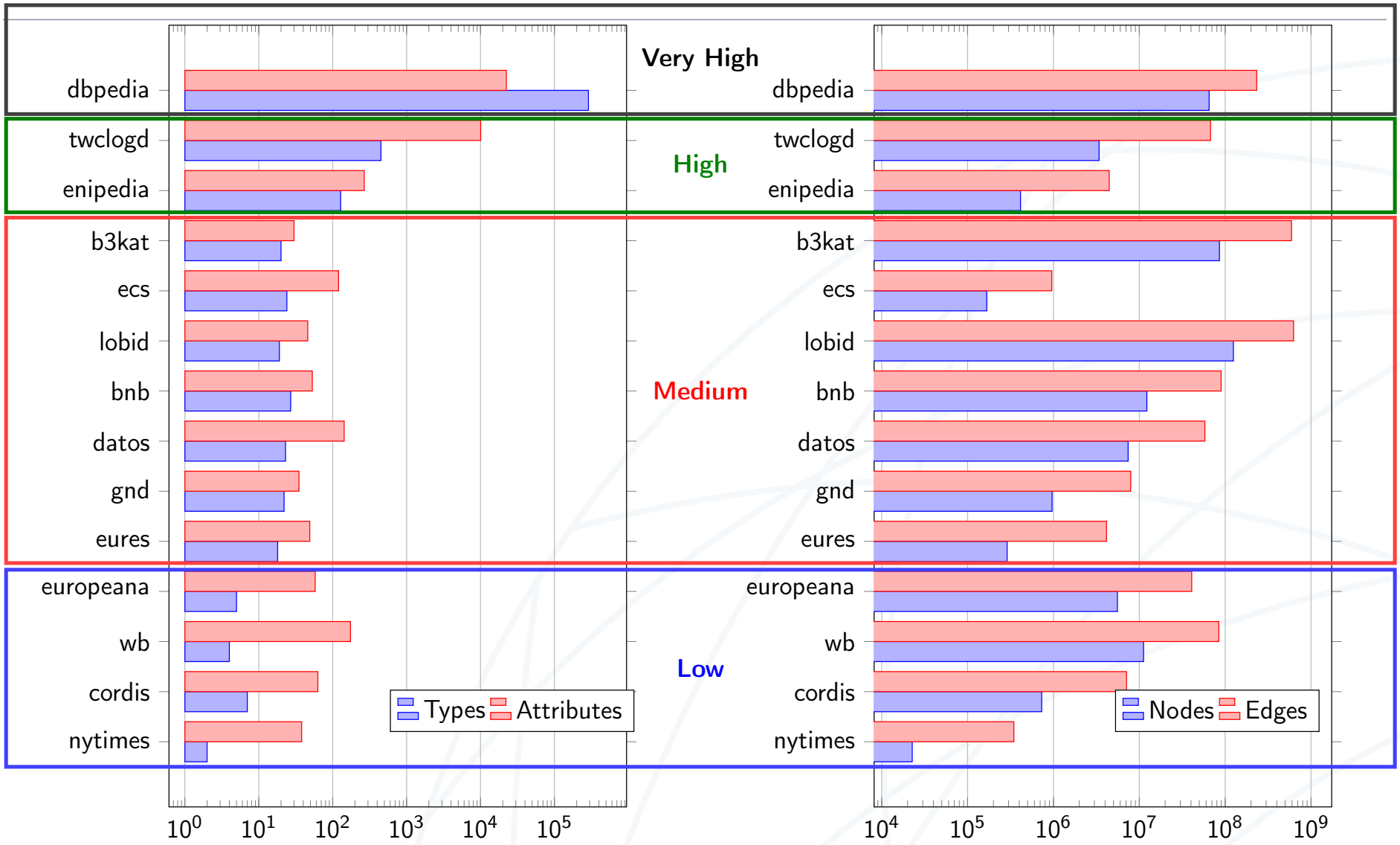
$$\frac{1}{|\mathcal{W}_{fbt}|} \times \sum_{c \in \mathcal{W}} \frac{1}{|C(c)|} \times \sum_{x \in C(c)} Prec(R, x)$$

Datasets

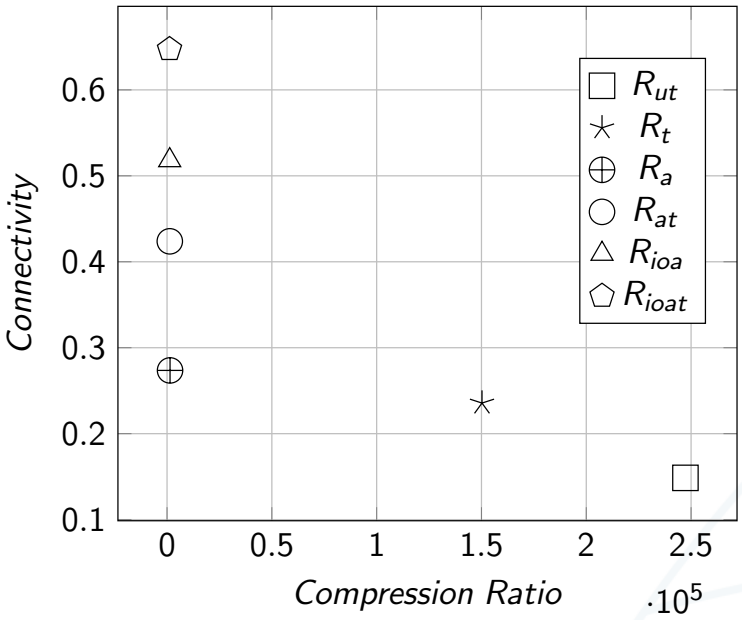

dbpedia
twclogd
enipedia
b3kat
ecs
lobid
bnb
datos
gnd
eures
europeana
wb
cordis
nytimes
Types
Attributes
10^0
10^1
10^2
10^3
10^4
10^5


dbpedia
twclogd
enipedia
b3kat
ecs
lobid
bnb
datos
gnd
eures
europeana
wb
cordis
nytimes
Nodes
Edges
10^4
10^5
10^6
10^7
10^8
10^9

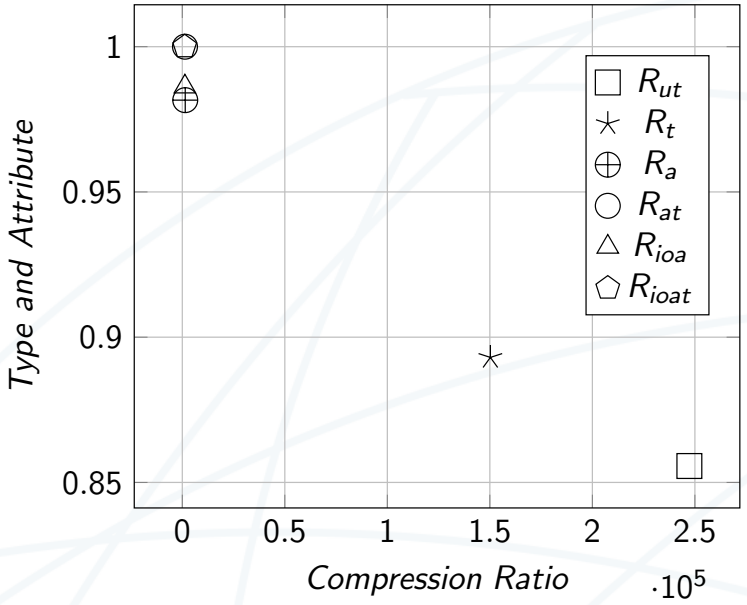
Datasets


Very High
High
Medium
Low
dbpedia
twclogd
enipedia
b3kat
ecs
lobid
bnb
datos
gnd
eures
europeana
wb
cordis
nytimes
Types
Attributes
Nodes
Edges
10^0 10^1 10^2 10^3 10^4 10^5
10^4 10^5 10^6 10^7 10^8 10^9

Tradeoffs: Precision VS. Compression

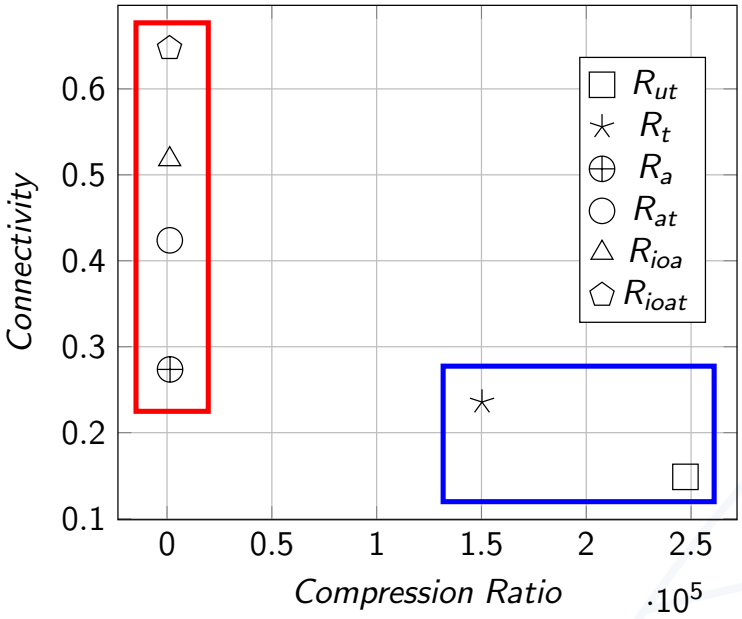


(a) Connectivity precision versus compression ratio.

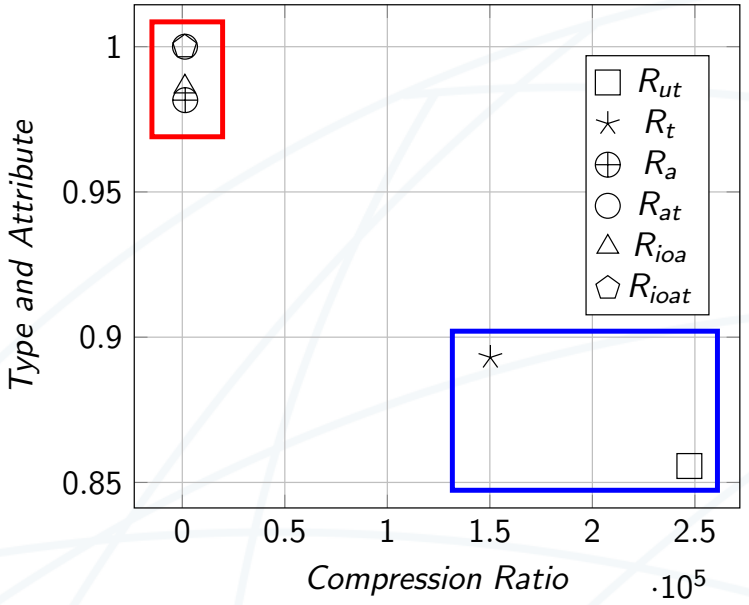


(b) Type and attribute precision versus compression ratio.

Tradeoffs: Precision VS. Compression

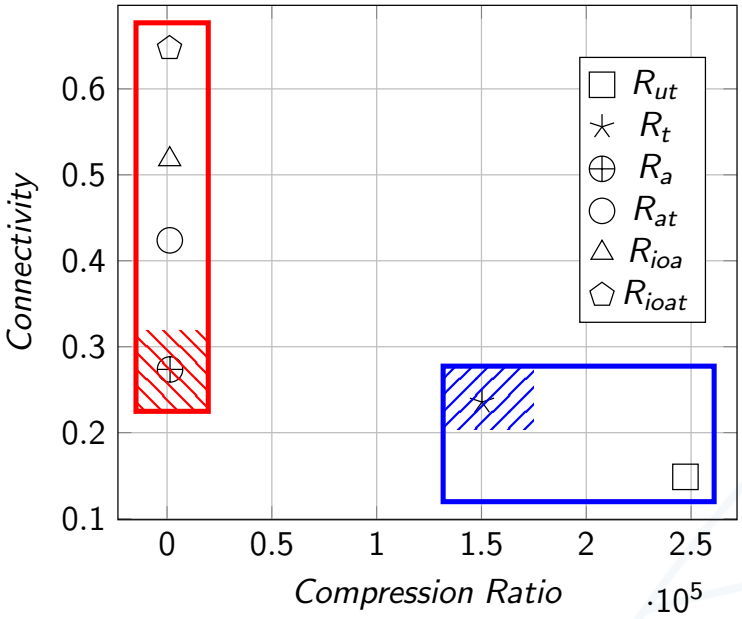


(a) Connectivity precision versus compression ratio.

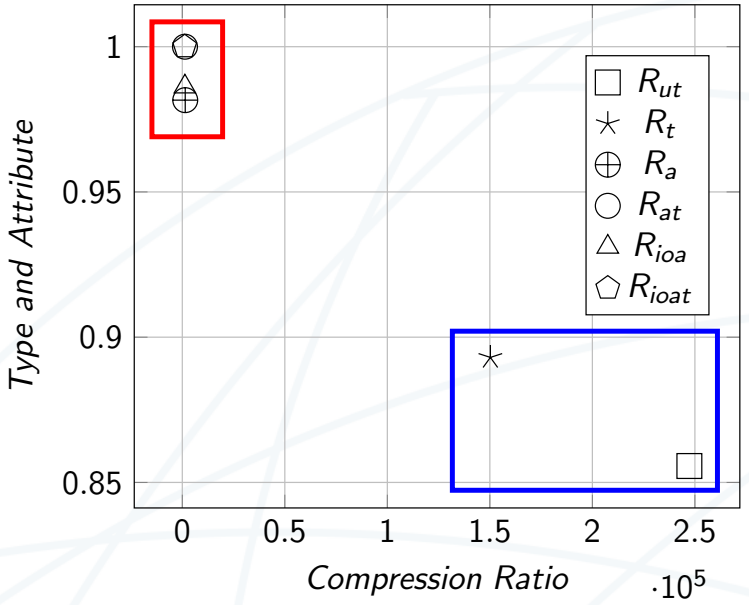


(b) Type and attribute precision versus compression ratio.

Tradeoffs: Precision VS. Compression

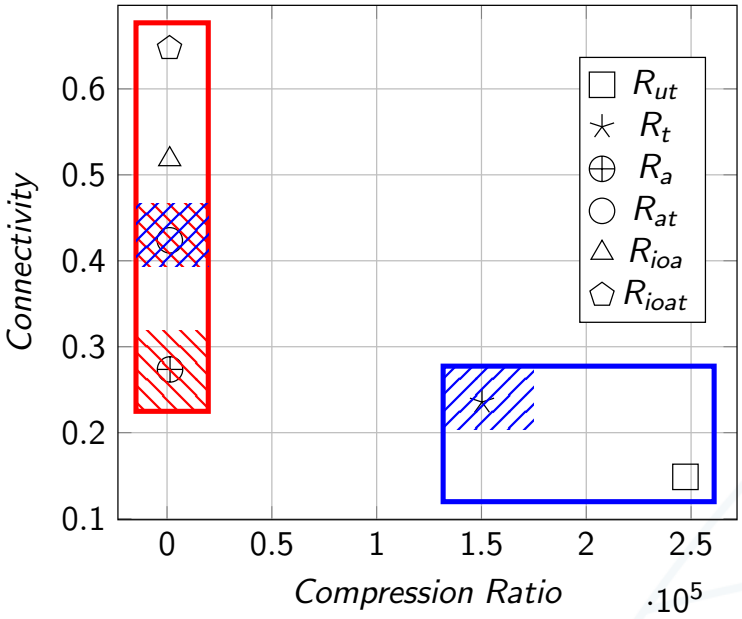


(a) Connectivity precision versus compression ratio.

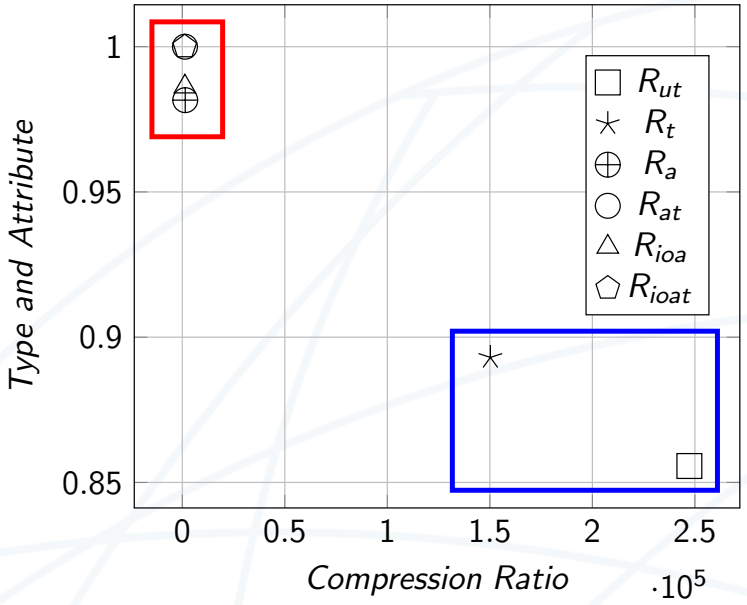


(b) Type and attribute precision versus compression ratio.

Tradeoffs: Precision VS. Compression

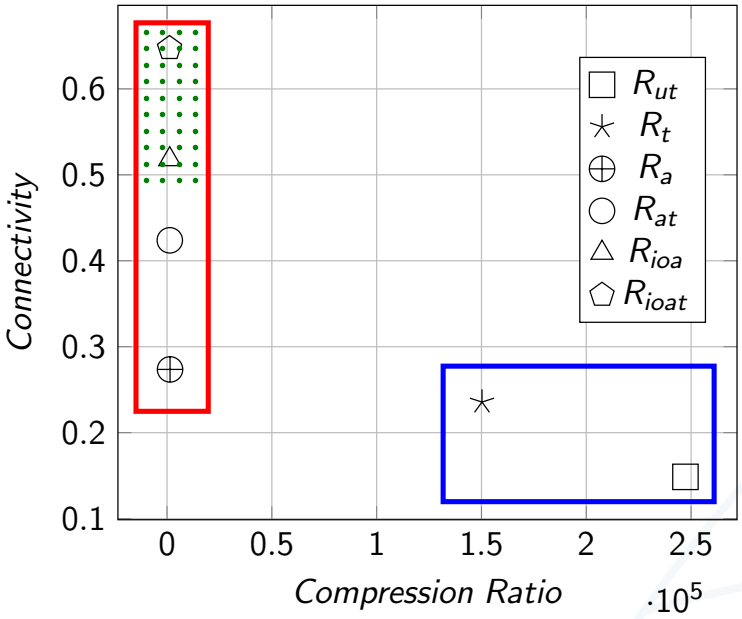


(a) Connectivity precision versus compression ratio.

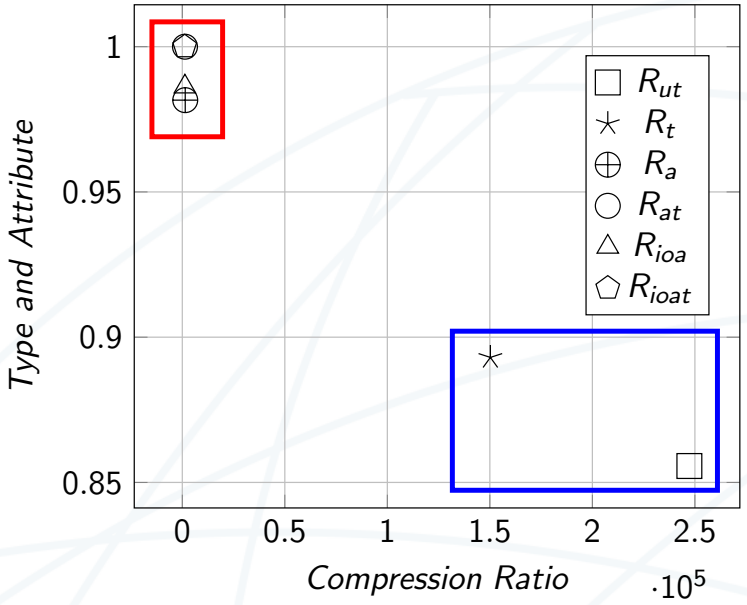


(b) Type and attribute precision versus compression ratio.

Tradeoffs: Precision VS. Compression

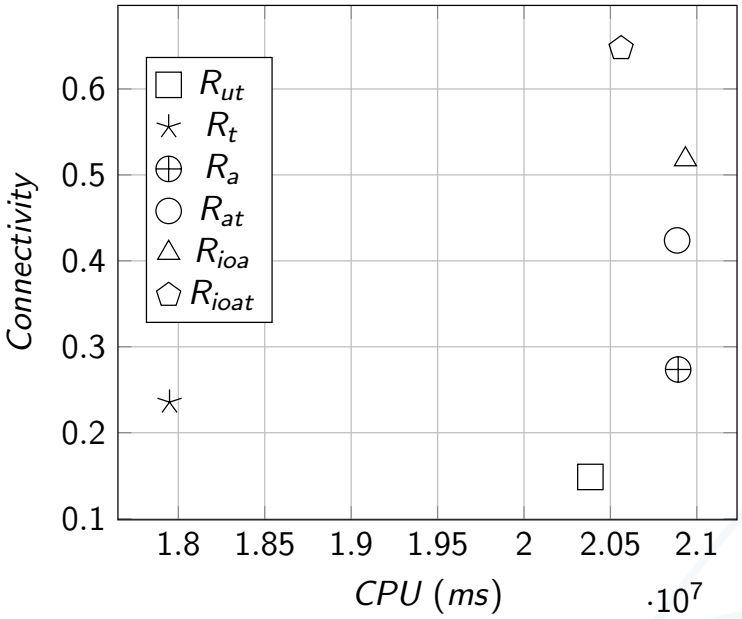


(a) Connectivity precision versus compression ratio.

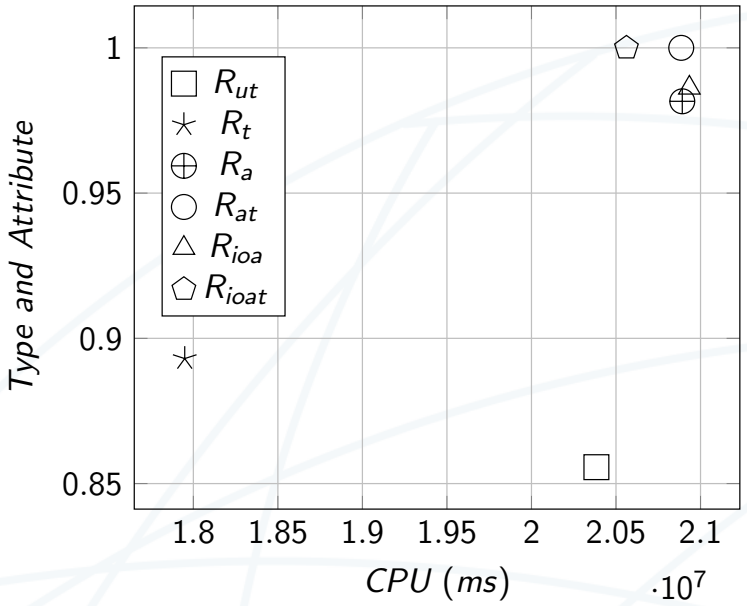


(b) Type and attribute precision versus compression ratio.

Tradeoffs: Precision VS. Performance

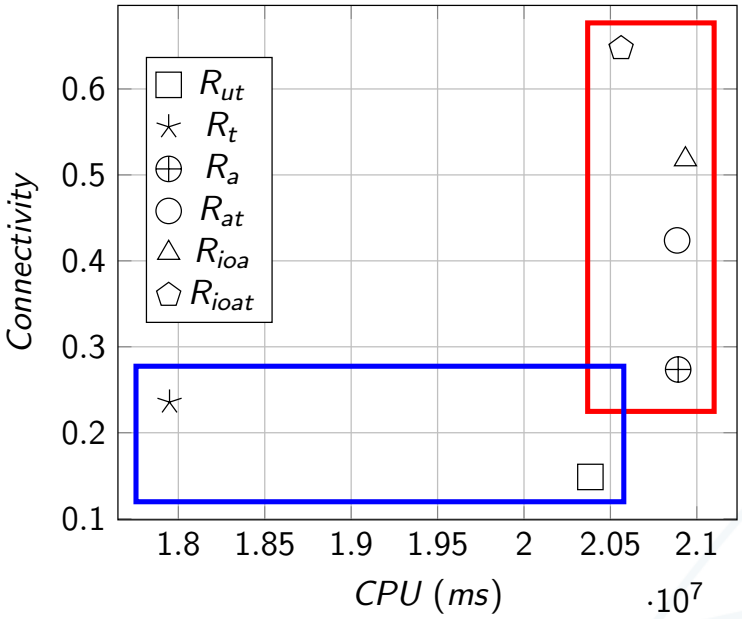


(a) Connectivity precision versus summarisation performance.

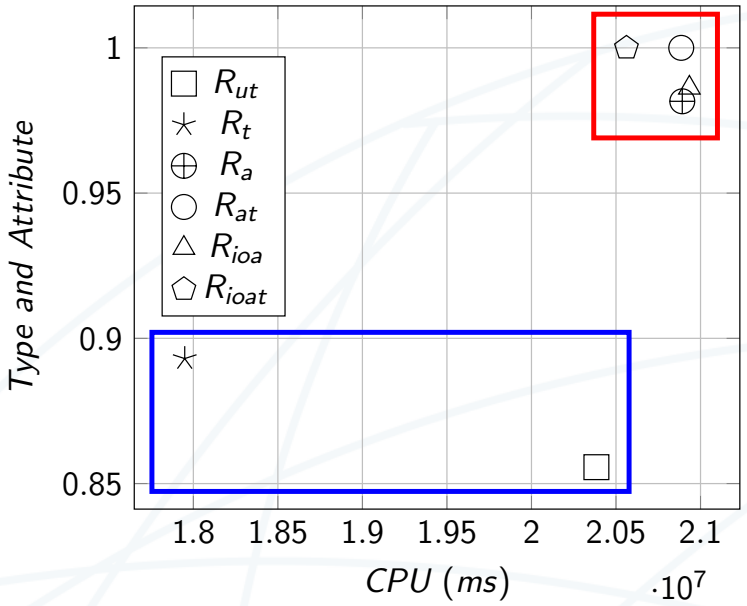


(b) Type and attribute precision versus summarisation performance.

Tradeoffs: Precision VS. Performance

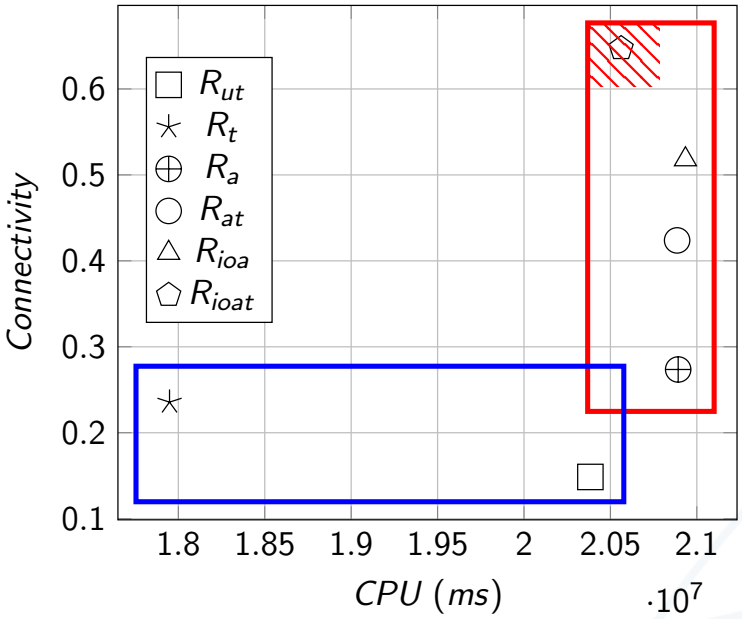


(a) Connectivity precision versus summarisation performance.

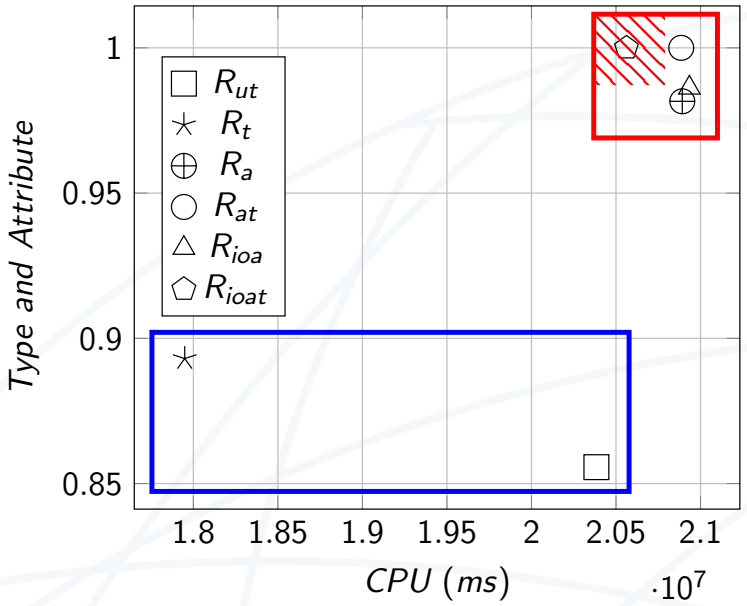


(b) Type and attribute precision versus summarisation performance.

Tradeoffs: Precision VS. Performance

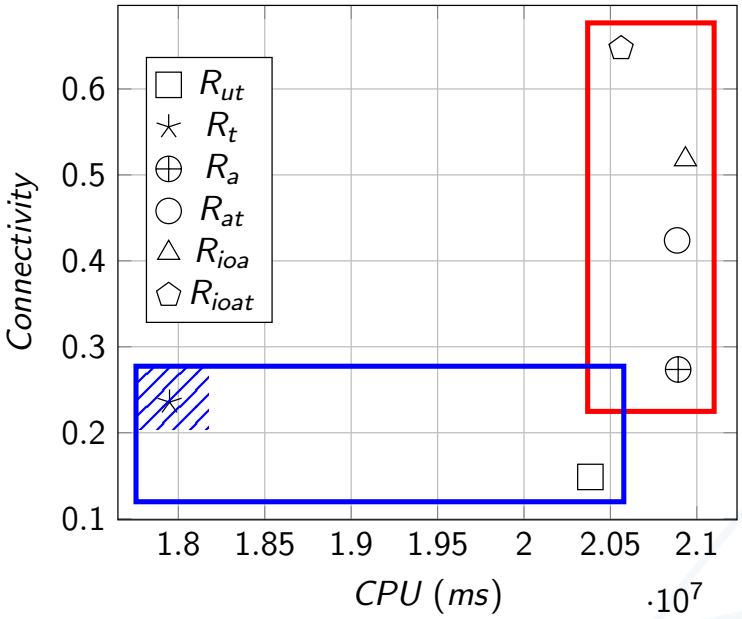


(a) Connectivity precision versus summarisation performance.

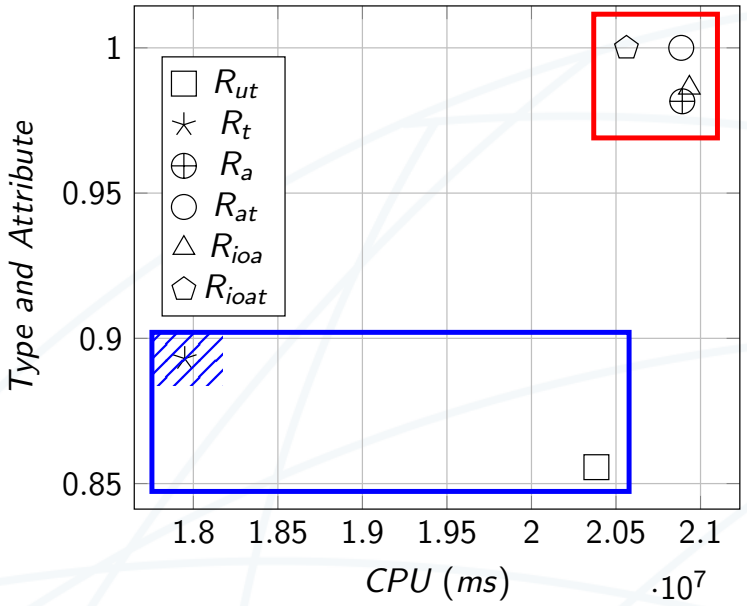


(b) Type and attribute precision versus summarisation performance.

Tradeoffs: Precision VS. Performance

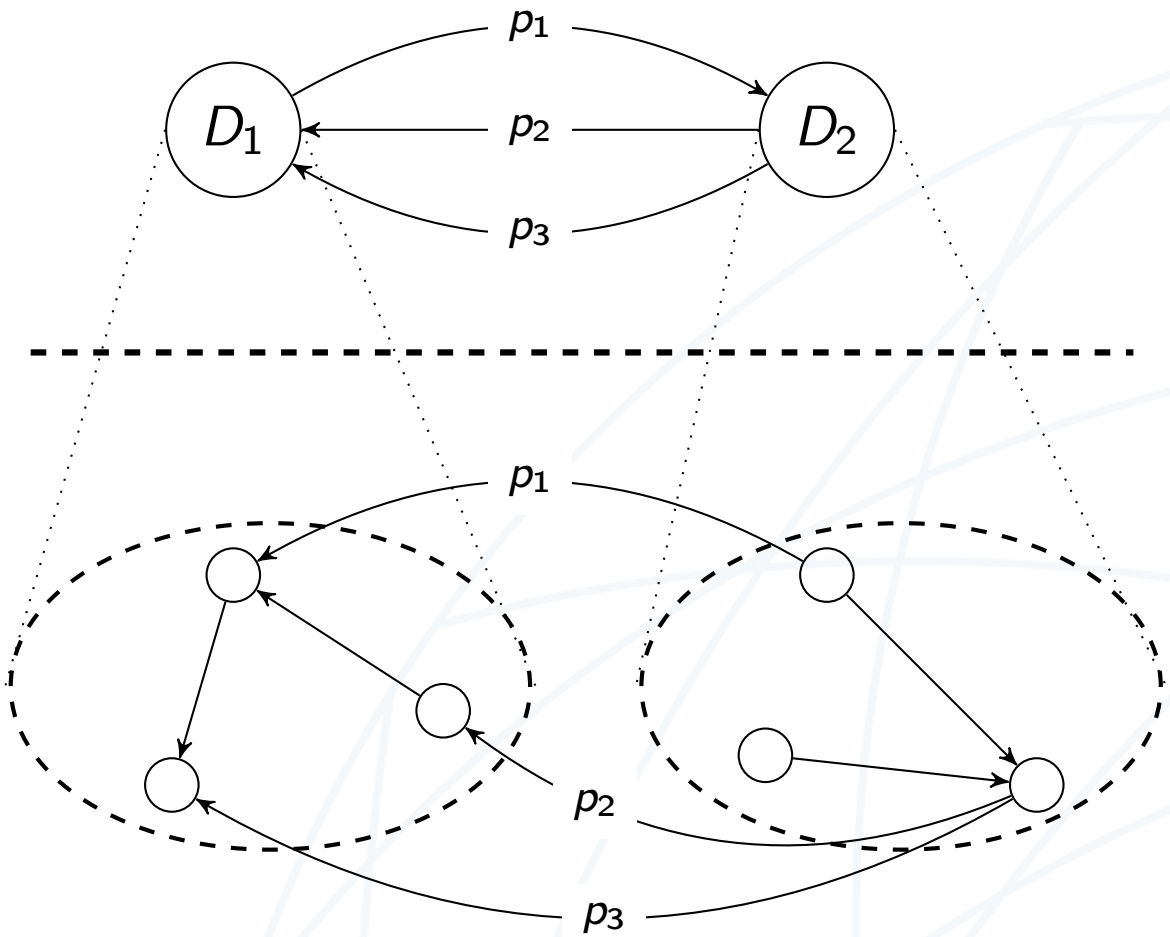


(a) Connectivity precision versus summarisation performance.



(b) Type and attribute precision versus summarisation performance.

Two-Layer Graph Data Model



Assisted SPARQL Query Editor

Application

- Provide suggestions for a missing element in a SPARQL query
- Possible **context-aware** suggestions:
  - Graph
  - Type
  - Attribute
  - Relation between two triple patterns

Solution

Map the SPARQL query to the RDF representation of the summary in order to get suggestions.

Example: Attribute Suggestions

```
1   ASK WHERE {
2     :article1 a :Article .
3
4
5
6
7     :article1 :title ?t .
8
9
10
11    ?i a :Institute .
12
13
14
15
16    ?i :employs ?p .
17
18
19
20    ?p :name "Renaud" .
21
22
23
24    ?p <                    # POF
25
26
27  }
```

Example: Attribute Suggestions

```
1   ASK WHERE {
2     :article1 a :Article .
3
4
5
6
7     :article1 :title ?t .
8
9
10
11    ?i a :Institute .
12
13
14
15
16    ?i :employs ?p .
17
18
19
20    ?p :name "Renaud" .
21
22
23
24    ?p <                    # POF
25
26
27  }
```

Example: Attribute Suggestions


```
1   ASK WHERE {
2     :article1 a :Article .
3
4
5
6
7     :article1 :title ?t .
8
9
10
11    ?i a :Institute .
12
13
14
15
16    ?i :employs ?p .
17
18
19
20    ?p :name "Renaud" .
21
22
23
24    ?p <                    # POF
25
26
27  }
```

```
2     ?n1 :feature [
3      :label :Article,
4      :type rdf:type
5     ] .
```

Example: Attribute Suggestions

```
1   ASK WHERE {
2     :article1 a :Article .
3
4
5
6
7     :article1 :title ?t .
8
9
10
11    ?i a :Institute .
12
13
14
15
16    ?i :employs ?p .
17
18
19
20    ?p :name "Renaud" .
21
22
23
24    ?p <                    # POF
25
26
27  }
```

```
2     ?n1 :feature [
3       :label :Article,
4       :type rdf:type
5     ] .
6
7     ?e1 :source ?n1 ;
8        :target _:b1 ;
9        :label :title .
```

Example: Attribute Suggestions

```
ASK WHERE {
  :article1 a :Article .

  :article1 :title ?t .

  ?i a :Institute .

  ?i :employs ?p .

  ?p :name "Renaud" .

  ?p <                    # POF

}
```

```
  ?n1 :feature [
   :label :Article,
   :type rdf:type
  ] .

  ?e1 :source ?n1 ;
      :target _:b1 ;
      :label :title .

  ?n2 :feature [
   :label :Institute,
   :type rdf:type
  ] .

  ?e2 :source ?n2 ;
      :target ?n3 ;
      :label ?employs .

  ?e3 :source ?n3 ;
      :target _:b2 ;
      :label :name .
```

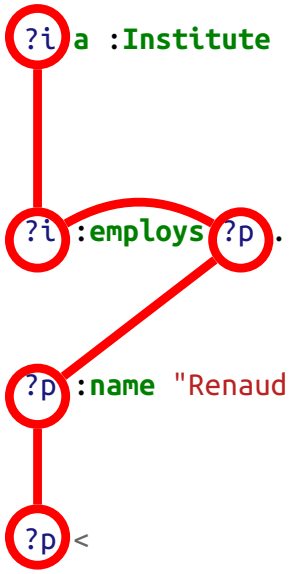
Example: Attribute Suggestions

```
1   ASK WHERE {
2     :article1 a :Article .
3
4
5
6
7     :article1 :title ?t .
8
9
10
11    ?i a :Institute .
12
13
14
15
16    ?i :employs ?p .
17
18
19
20    ?p :name "Renaud" .
21
22
23
24    ?p <                    # POF
25
26
27  }
```

```
1   SELECT ?POF {
2     ?n1 :feature [
3      :label :Article,
4      :type rdf:type
5     ] .
6
7     ?e1 :source ?n1 ;
8        :target _:b1 ;
9        :label :title .
10
11    ?n2 :feature [
12     :label :Institute,
13     :type rdf:type
14    ] .
15
16    ?e2 :source ?n2 ;
17       :target ?n3 ;
18       :label ?employs .
19
20    ?e3 :source ?n3 ;
21       :target _:b2 ;
22       :label :name .
23
24    ?e4 :source ?n3 ;
25       :target _:b3 ;
26       :label ?POF .
27  }
```

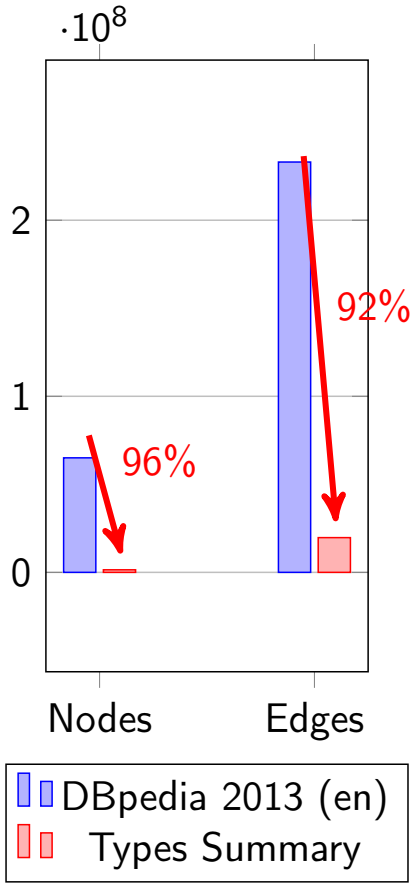
Example: Attribute Suggestions

```
ASK WHERE {
  :article1 a :Article .

  :article1 :title ?t .

  ?i a :Institute .

  ?i :employs ?p .

  ?p :name "Renaud" .

  ?p <                          # POF

}
```



```
SELECT ?POF {

    ?n2 :feature [
      :label :Institute,
      :type rdf:type
    ] .

    ?e2 :source ?n2 ;
        :target ?n3 ;
        :label ?employs .

    ?e3 :source ?n3 ;
        :target _:b2 ;
        :label :name .

    ?e4 :source ?n3 ;
        :target _:b3 ;
        :label ?POF .
}
```

Summary And Directions for Future Research

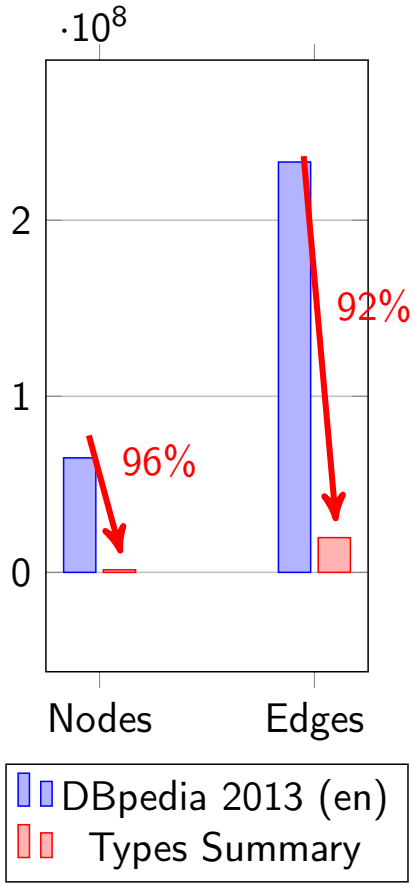


Summary Updates: Efficiently updates the summary of a dataset without re-doing the process from scratch.

Data Quality: Improve the data modeling by identifying structural inconsistencies.

Approximate Summaries: Develop summarisation relations which have little impact on the connectivity precision, and yet generate small summaries.

Summary And Directions for Future Research

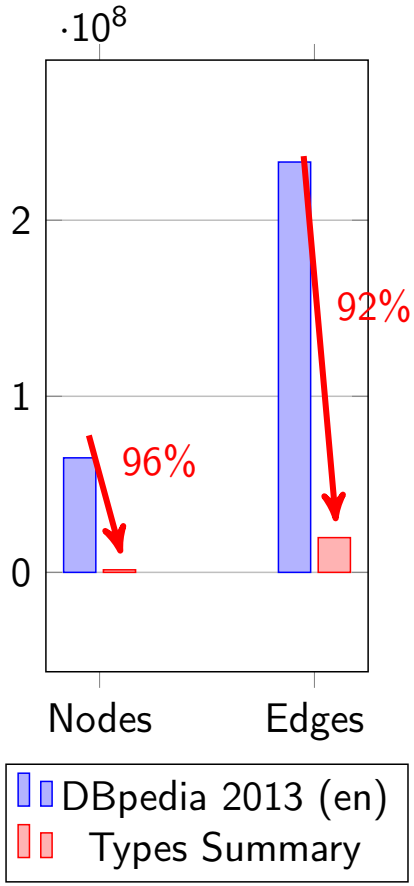


Summary Updates: Efficiently updates the summary of a dataset without re-doing the process from scratch.

Data Quality: Improve the data modeling by identifying structural inconsistencies.

Approximate Summaries: Develop summarisation relations which have little impact on the connectivity precision, and yet generate small summaries.

Summary And Directions for Future Research

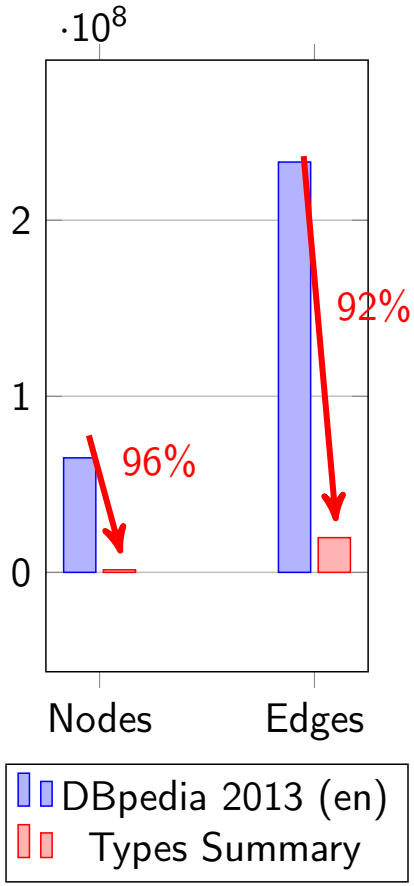


Summary Updates: Efficiently updates the summary of a dataset without re-doing the process from scratch.

Data Quality: Improve the data modeling by identifying structural inconsistencies.

Approximate Summaries: Develop summarisation relations which have little impact on the connectivity precision, and yet generate small summaries.

Summary And Directions for Future Research



Summary Updates: Efficiently updates the summary of a dataset without re-doing the process from scratch.

Data Quality: Improve the data modeling by identifying structural inconsistencies.

Approximate Summaries: Develop summarisation relations which have little impact on the connectivity precision, and yet generate small summaries.

Contributions

Graph Summarization:

- **Introducing RDF Graph Summary with application to Assisted SPARQL Formulation**, *WebS Workshop, DEXA2012*
- **Efficiency and precision trade-offs in graph summary algorithms**, *Proceedings of the 17th International Database Engineering Applications Symposium*
- **Live SPARQL auto-completion**, *Proceedings of the 2014 International Conference on Posters Demonstrations Track*

Ranking:

- **The Sindice-2011 dataset for entity-oriented search in the web of data**, *1st international workshop on entity-oriented search (EOS)*
- **Sindice BM25F at SemSearch 2011**, *Proceedings of the 4th International Semantic Search Workshop*
- **Effective retrieval model for entity with multi-valued attributes: Bm25mf and beyond**, *International Conference on Knowledge Engineering and Knowledge Management*

Graph Ranking