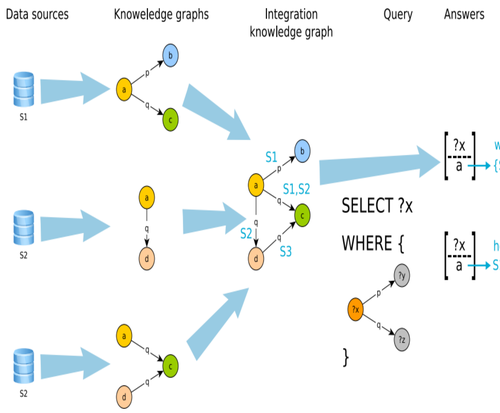
NPCS: Native Provenance Computation for SPARQL

Zubaria Asma[1,2], Daniel Hernández[3], Luis Galárraga[4], Giorgos Flouris[1], Irini Fundulaki[1] and Katja Hose[5]

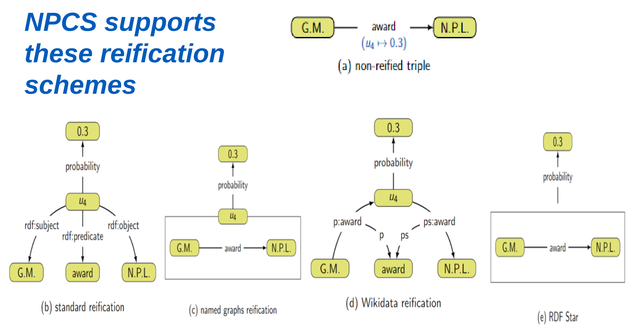
{[1]FORTH-ICS,[2]University of Crete} Heraklion, Crete, Greece, [3]University of Stuttgart, Stuttgart, Germany ,[4]Inria, Rennes, France,[5]TU Wien, Wien, Austria

Introduction

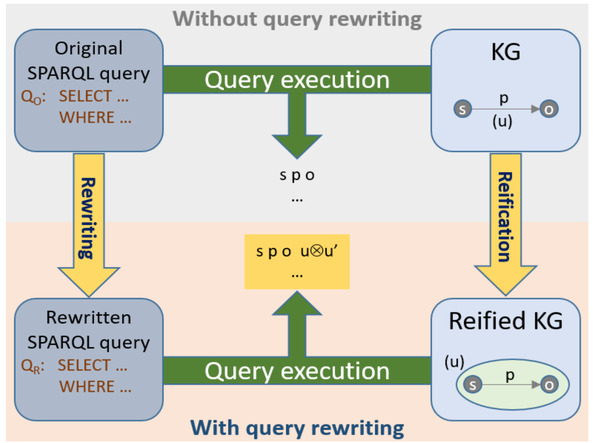
- Knowledge Graphs (KGs)
  - Important in both academia and industry
  - Ideal for integrating data from various sources
- Provenance
  - Critical for trust assessment and dynamic data
- Our Contribution (NPCS)
  - Enriches results with how-provenance annotations
  - Supports monotonic and non-monotonic SPARQL
  - NPCS rewrites SPARQL query Q into Q', generating how-provenance polynomials in the spm–semiring K = (K, $\oplus$, $\otimes$, $\ominus$, 0, 1) for K-graph G.



NPCS supports these reification schemes



Schema for NPCS architecture



Query: Get all women awarded with nobel prize in Literature

[5]

SPARQL Query

```
SELECT ?x
WHERE {
?x gender female .
?x award "N.P.Literature"
}
```

Result

?x
G. Mistral
O. Tokarczuk

NPCS Query

```
(SELECT ?x AggSum(?prov)) AS ?k)
WHERE {
Reify(?x , gender ,female , ?prov0) .
Reify(?x , award ,"N.P.Literature", ?prov1)
BIND(Prod (?prov0, ?prov1) as ?prov)
} group by(?x))
```

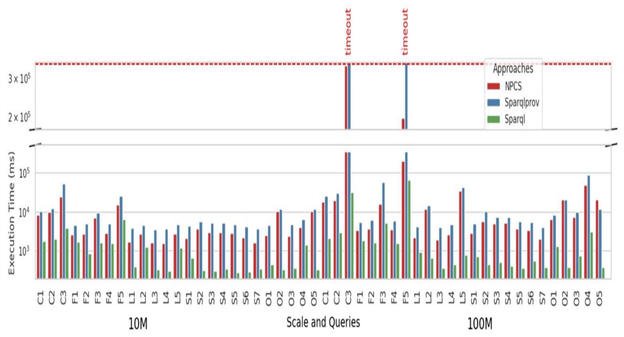
NPCS Result

?x	k
G. Mistral	$k_1 \otimes k_3$
O. Tokarczuk	$k_4 \otimes k_7$

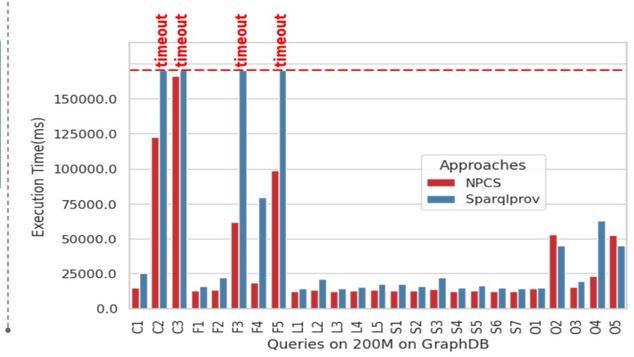
Evaluation of NPCS

- Tested on different reification schemes using two engines (Stardog, GraphDB) with datasets of 10M, 100M, 200M from WatDiv and 15 billion triples from Wikidata
- Observed trends: NPCS consistently outperforms SPARQLprov in 48 out of 50 studied cases
- Applicable to any standard RDF/SPARQL engine with a significant performance margin

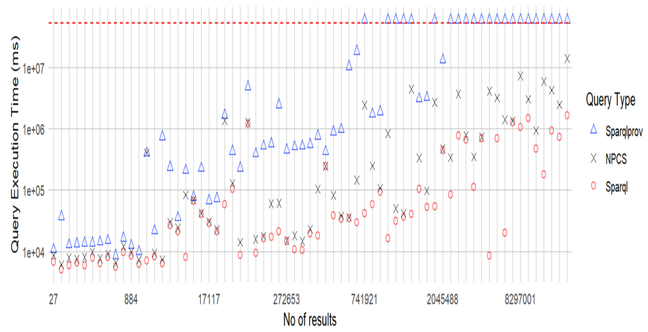
NPCS performs better on 10M and 100M datasets



On 200M, NPCS exhibits a more significant performance, with SPARQLprov experiencing increased timeouts



Graph showing results vs. execution time for WDBench queries on Wikidata



Conclusion

- Novel SPARQL-based method for computing how-provenance annotations
- Outperforms
- existing solutions
- Enables efficient computation of provenance for millions of query results on large knowledge graphs
- Ideal for ETL processes in multi-source KG construction

Acknowledgment

This work was supported by the ITN KnowGraphs project, under the EU H2020 Marie Skłodowska-Curie grant agreement No 860801

References

1. Daniel Hernández, Luis Galárraga, and Katja Hose. 2021. Computing How-Provenance for SPARQL Queries via Query Rewriting. Proceedings of the VLDB Endowment 14, 13 (2021), 3389–340.
2. Floris Geerts, Thomas Unger, Grigoris Karvounarakis, Irini Fundulaki, and Vassilis Christophides. 2016. Algebraic Structures for Capturing the Provenance of SPARQL Queries. Journal of the ACM 63, 1 (2016), 7:1–7:63.

FORTH FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS · ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE · University of Stuttgart Germany · Inria La Fondation · TU WIEN TECHNISCHE UNIVERSITÄT WIEN Vienna | Austria