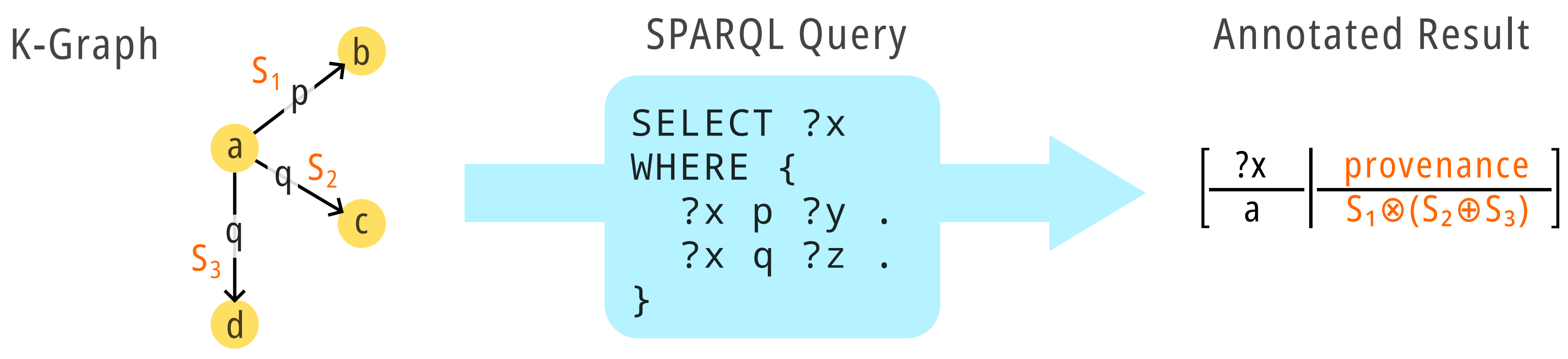
NPCS: Native Provenance Computation for SPARQL

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, Greece, [3]University of Stuttgart, Germany , [4]Inria, Rennes, France, [5]TU Wien, Austria

Knowledge Graphs: Data model 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 a K-graph G.

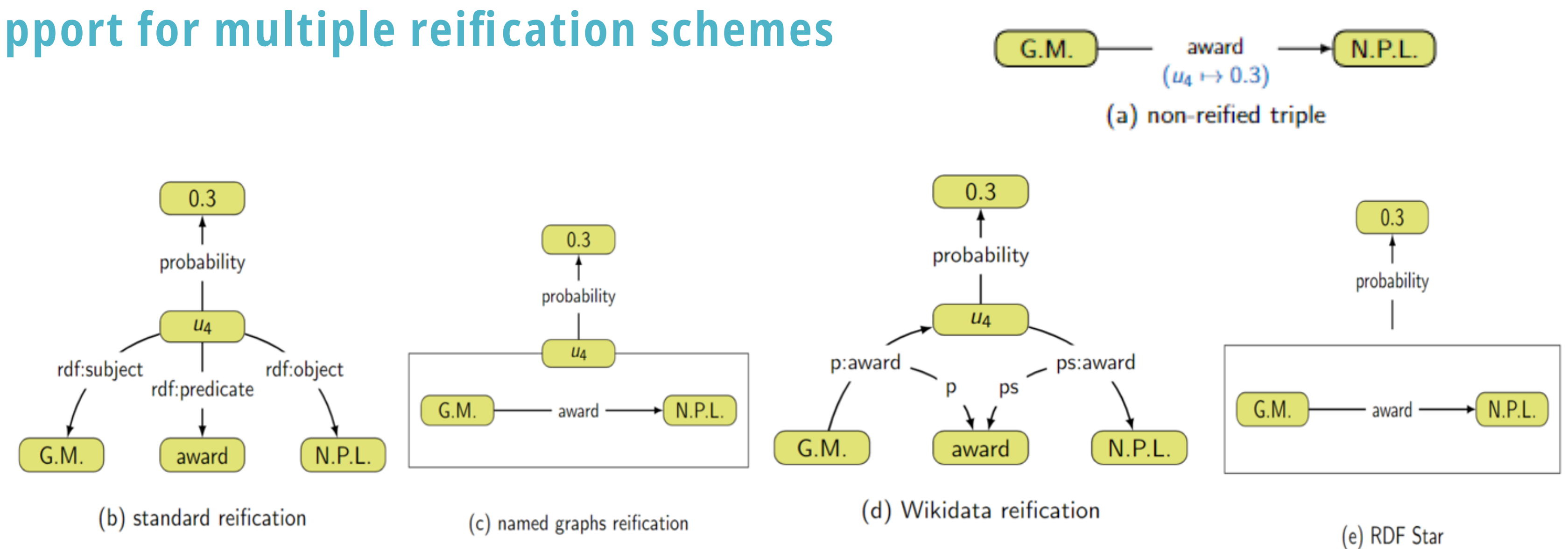
K-Annotated SPARQL Algebra [Geerts et al.]



```
SELECT ?x
WHERE {
  ?x p ?y .
  ?x q ?z .
}
```

?x	provenance
a	$S_1 \otimes (S_2 \oplus S_3)$

Support for multiple reification schemes



(a) non-reified triple

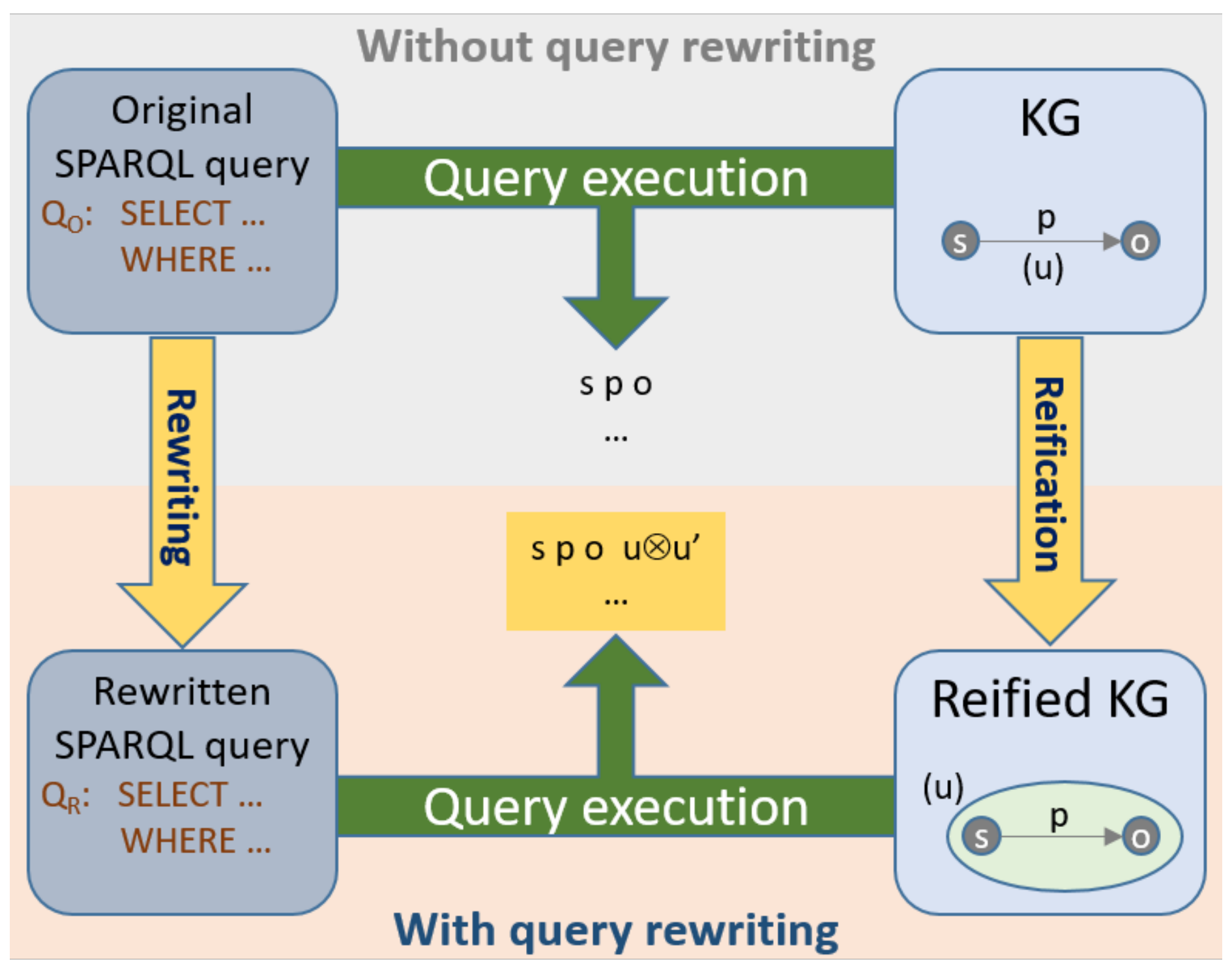
(b) standard reification

(c) named graphs reification

(d) Wikidata reification

(e) RDF Star

NPCS Architecture



Without query rewriting

Original SPARQL query
Q_O: SELECT … WHERE …

Query execution

KG

s p o (u)

s p o
…

Rewriting

Reification

s p o u⊗u'
…

Rewritten SPARQL query
Q_R: SELECT … WHERE …

Query execution

Reified KG

(u) s p o

With query rewriting

Get all awarded women

SPARQL Query

```
SELECT DISTINCT  ?x
WHERE  {
   ?x  gender female .
   ?x  award ?y .
}
```

Result

?x
G.Mistral
O.Tokarczuk

SPARQLprov Query [Hernandez et al.]

```
SELECT  ?x ?k⊕⊗0 ?k⊕⊗1 ?k⊕
WHERE  {
   Reify(?x, gender, female, ?k⊕⊗0) .
   Reify(?x, award, ?y, ?k⊕⊗1) .
   BIND(B(?x) as ?k⊕)
}
```

SPARQLprov Result

?x	?k⊕⊗0	?k⊕⊗0	?k⊕
G.Mistral	S_1	S_2	B_1
G.Mistral	S_1	S_3	B_1
O.Tokarczuk	S_4	S_5	B_2

NPCS Query

```
SELECT  ?x (AggSum(?k⊕) as ?k)
WHERE  {
   Reify(?x, gender, female, ?k⊕⊗0) .
   Reify(?x, award, ?y, ?k⊕⊗1) .
   BIND(Prod(?k⊕⊗0, ?k⊕⊗1) as ?k⊕)
}
GROUP BY ?x
```

NPCS Result

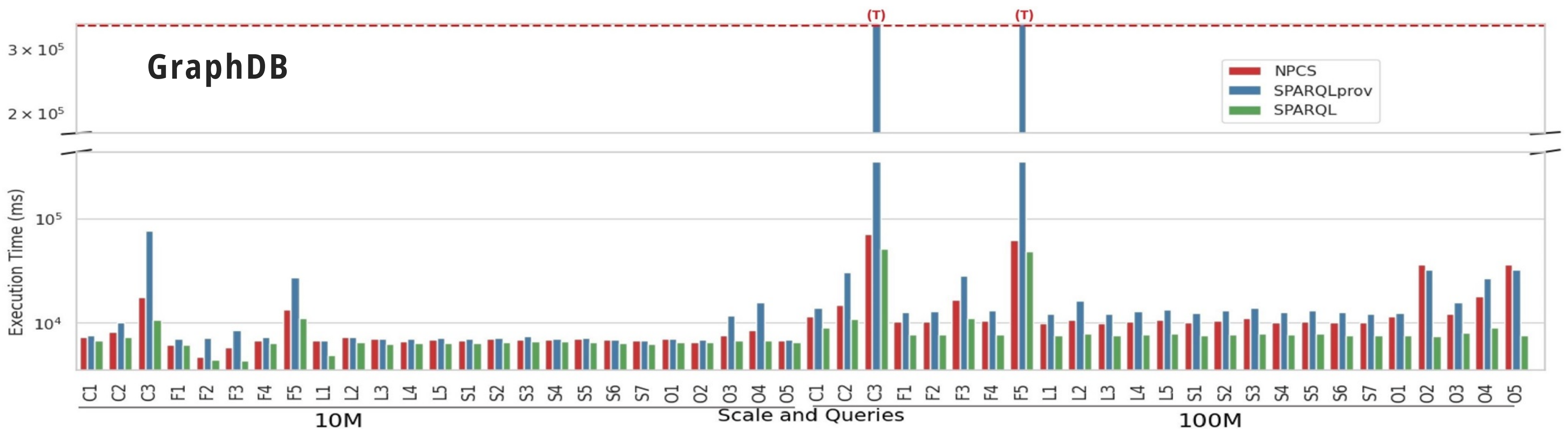
?x	?k
G.Mistral	$(S_1 \otimes S_2) \oplus (S_1 \otimes S_3)$
O.Tokarczuk	$(S_4 \otimes S_5)$

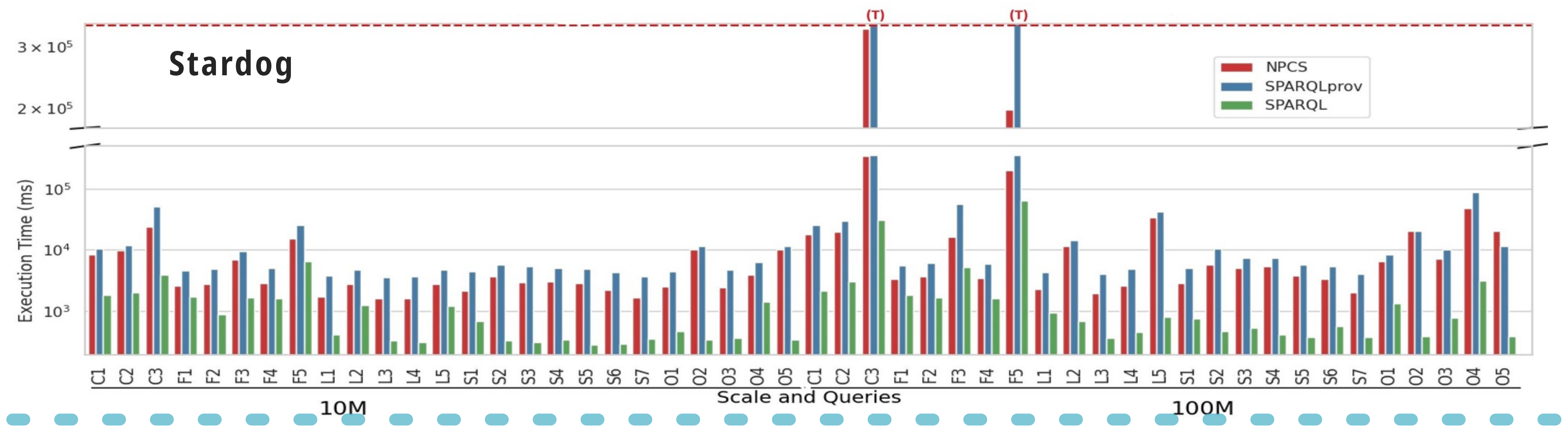
Evaluation of NPCS

Tested on different reification schemes using two engines (Stardog, GraphDB) with datasets of 10M, 100M, 200M from WatDiv and 15 billion triples from Wikidata.

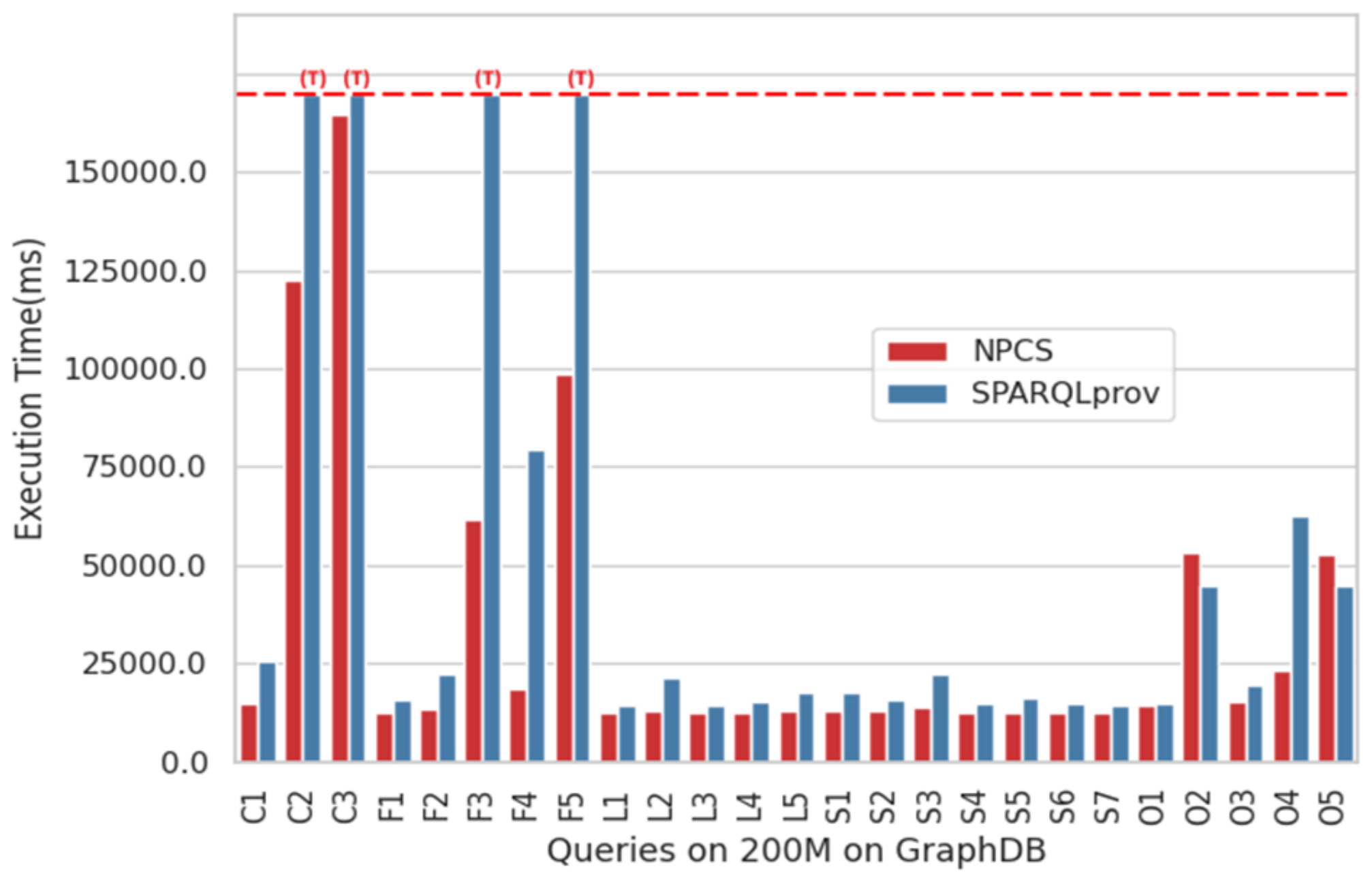
Observed trends: NPCS consistently outperforms SPARQLprov in 48 out of 50 studied cases
Applicable to any standard RDF/SPARQL engine with a significant performance margin

NPCS performs better on 10M and 100M Watdiv datasets

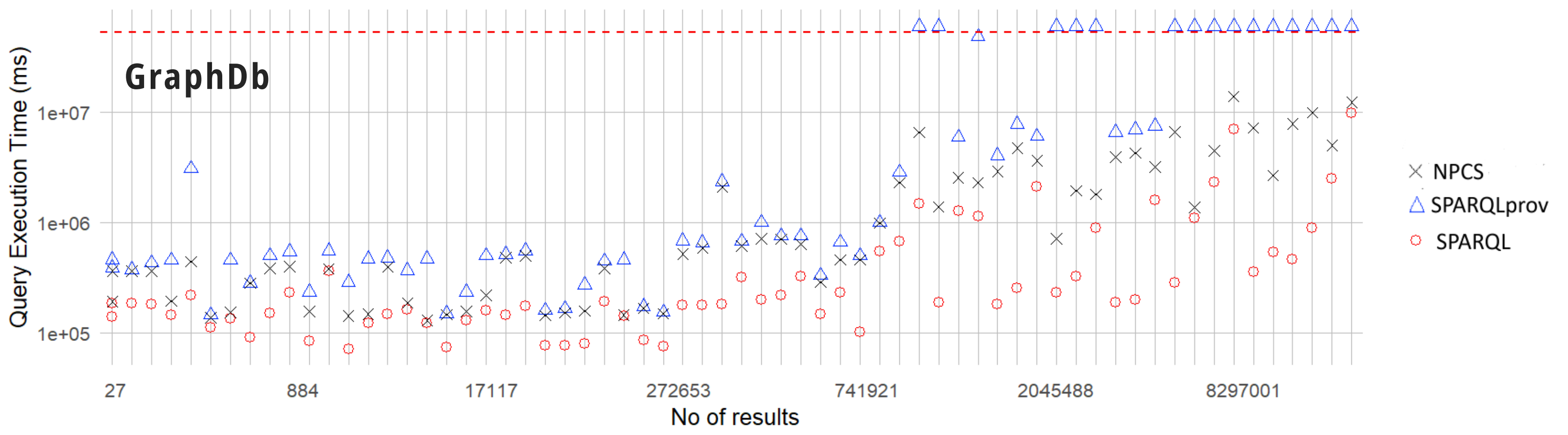


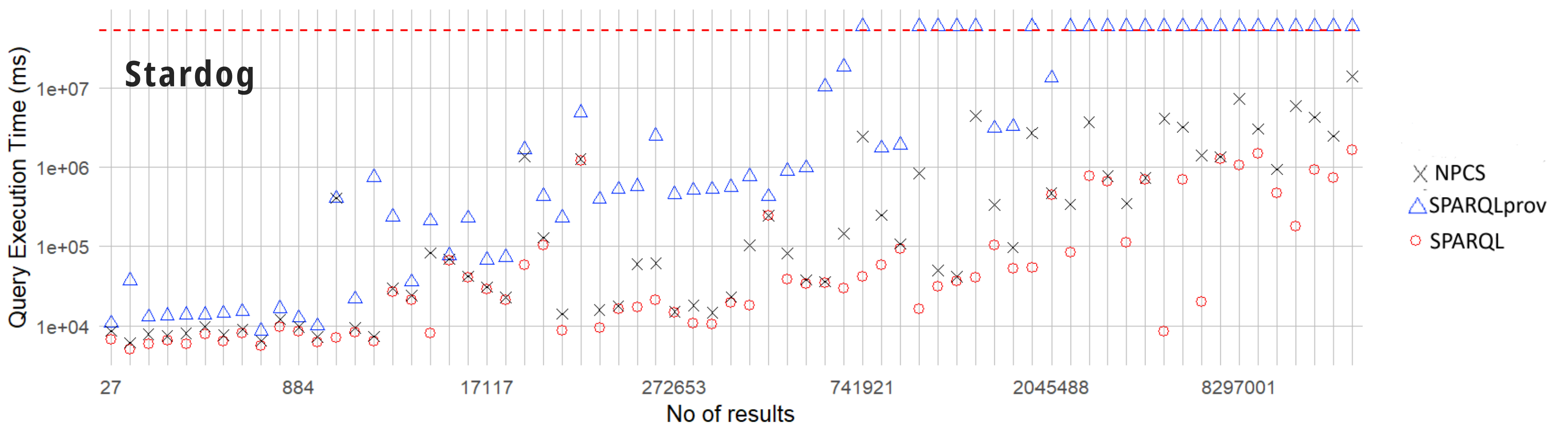


On 200M, NPCS exhibits a more significant performance, with SPARQLprov experienced increased timeouts



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

This work was partially funded by the EU H2020 R&I Marie Skłodowska-Curie program, project KnowGraphs – 860801; the TAILOR Network (EU Horizon 2020 R&I program under GA 952215); the COST Action CA19134; the Deutsche Forschungsgemeinschaft (DFG): Germany's Excellence Strategy–EXC 2120/1 – 390831618; and the DFG: SPP 1921 – 318363223 (COFFEE STA 572_15-2).

1. F. Geerts, T. Unger, G. Karvounarakis, I. Fundulaki, and V.s Christophides. 2016. Algebraic Structures for Capturing the Provenance of SPARQL Queries. Journal of the ACM 63, 1 (2016), 7:1–7:63.
2. D. Hernández, L. Galárraga, and K. Hose. 2021. Computing How-Provenance for SPARQL Queries via Query Rewriting. Proceedings of the VLDB Endowment 14, 13 (2021), 3389–340.