Distributed SPARQL Throughput Increase: On the Effectiveness of Workload driven RDF partitioning

Problem:
- One RDF store not enough ➤ partitioning needed
- Distributed joins dominate simple queries

Proposed solution
- Partition data given workload sample
- up to 8.5x better than random part.
- up to 5.2x better than horizontal part.

Process:
1. Log SPARQL workload & results
2. Partition Queries Graph (Metis)
3. Replication (triples on border)
4. Placement (triples ➤ partitions)

Performance metric:
# of distributed joins

Results:
how many queries do I need to log?

Performance comparison: traditional vs. workload-driven partitioning

Cosmin Basca
Abraham Bernstein

www.ifi.uzh.ch/ddis/