Distributed Query Evaluation Over Constrained Database

Avi Tomar [1]

[1] Vrije Universiteit Brussel
avi.tomar@vub.be

Join queries are a fundamental operation in large-scale data processing. Since processing these joins at scale requires distributing data across a cluster, the massively parallel computation (MPC) [1] model has become the standard theoretical framework for reasoning about the cost of algorithms running in a shared-nothing system. In this model, performance is not measured by computation but by communication: the maximum load received by any single server in any of the rounds and the number of synchronization rounds required. Thus, optimizing join evaluation in such systems then amounts to optimizing the communication cost incurred by the join algorithm. It is common that algorithms in the MPC model are restricted to either single round or constant round.

Existing results largely assume arbitrary input relations. In practice, database instances are far from arbitrary; they exhibit structural regularities given by schema-level constraints such as primary keys, functional dependencies, and bounded degrees of attributes [2]. Existing work on MPC join algorithms neither exploits this structure nor characterizes how such constraints can be leveraged to achieve better load. This leaves open a concrete question: how do schema-level constraints affect the communication complexity of computing a join query in single and multiple rounds in the MPC model, and when do they admit provably cheaper algorithms than the unconstrained case?

In our work, we propose a single-round algorithm for join queries considering simple degree constraints i.e., functional dependencies. The load of our algorithm does not exceed the load guarantees of the worst-case optimal algorithm for the single-round setting which does not consider functional dependencies [3]. Beyond the single-round setting, We also present a multi-round algorithm for join queries considering functional dependencies.

References

[1] Paul Beame, Paraschos Koutris, and Dan Suciu. Skew in parallel query processing. PODS '14, page 212–223, New York, NY, USA, 2014. Association for Computing Machinery.

[2] Sungjin Im, Benjamin Moseley, Hung Ngo, and Kirk Pruhs. Efficient algorithms for cardinality estimation and conjunctive query evaluation with simple degree constraints. *Proc. ACM Manag. Data*, 3(2), June 2025.

[3] Paris Koutris, Semih Salihoglu, and Dan Suciu. Algorithmic aspects of parallel query processing. SIGMOD '18, page 1659–1664, New York, NY, USA, 2018. Association for Computing Machinery.