Technical Perspective: Structure and Complexity of Bag Consistency

Mihalis Yannakakis
Columbia University, USA

The paper *Structure and Complexity of Bag Consistency* by Albert Atserias and Phokion Kolaitis [1] studies fundamental structural and algorithmic questions on the global consistency of databases in the context of bag semantics. A collection D of relations is called *globally consistent* if there is a (so-called "universal") relation over all the attributes that appear in all the relations of D whose projections yield the relations in D. The basic algorithmic problem for consistency is: given a database D, determine whether D is globally consistent. An obvious necessary condition for global consistency is *local* (or *pairwise*) *consistency*: every pair of relations in D must be consistent. This condition is not sufficient however: It is possible that every pair is consistent, but there is no single global relation over all the attributes whose projections yield the relations in D. A natural structural question is: Which database schemas have the property that every locally consistent database over the schema is also globally consistent?

These questions were studied early on, as the theory of relational databases was being developed, in a model where relations are assumed to be sets, i.e. have no duplicate tuples. On the structural side, it was shown that the class of acyclic database schemas characterizes precisely those schemas for which local consistency is equivalent to global consistency [2]. On the algorithmic side, the global consistency problem was shown to be in general NP-complete [5]. If the database schema is acyclic however, the consistency problem can be solved in polynomial time.

The paper [1] by Atserias and Kolaitis addresses these basic questions in the model where the relations are bags (multisets). The relationship between the set-based and the bag-based model is by no means straightforward, and results may not necessarily carry over from one model to the other. For example, the classical problem of containment of conjunctive queries has been long well-understood in the set model (the problem is NP-complete [3]). However, the problem appears to be much harder in the bag model, and it is still not even known to be decidable; this discrepancy was pointed out in [4], which raised the issue of revisiting fundamental problems of database theory in the bag model.

Regarding the structural question of when local consistency is equivalent to global consistency, [1] shows that the same characterization holds in the bag model, i.e. the local-to-global consistency property holds precisely for the class of acyclic schemas. The proof however requires several new ideas and techniques (including linear programming and network flows) because some basic features of the set model do not hold any more in the bag model. For example, in the set model, if the database is consistent, then the join of all the relations can serve as the universal relation (i.e. its projections yield the relations of the database). This does not hold in the bag model even for two relations.

Regarding the algorithmic question of determining whether a given database is consistent, [1] provides a sharp dichotomy theorem for all fixed database schemas in the bag model: if the schema is acyclic then the problem can be solved in polynomial time, whereas if it is cyclic then the problem is NP-complete. The NP-completeness holds for example even for a schema with 3 relations and 3 attributes. Note, by contrast, that in the set model, if the schema is fixed, then the consistency problem can be solved trivially in polynomial time (even if the schema is cyclic), where the degree of the polynomial depends on the schema.

The paper addresses fundamental consistency questions in the bag model and resolves them exactly using nontrivial, elegant techniques. The relationship between local and global consistency arises in a variety of other settings, for example for probability distributions, and in quantum information. In the last section of their paper, Atserias and Kolaitis discuss a general model of relations over semirings, which includes as special cases the set and the bag model, and provide a preview of interesting forthcoming work on the local-to-global consistency problem in this general setting.

1. REFERENCES

[1] A. Atserias, P. Kolaitis. Structure and complexity of bag consistency. In *Proc. 40th ACM Symp. on Principles of Database Systems*, pp. 247-259, 2021.

[2] C. Beeri, R. Fagin, D. Maier, M. Yannakakis. On the desirability of acyclic database schemes. *Journal of ACM*, 30(3): 479-513, 1983.

[3] A. K. Chandra, P. Merlin. Optimal implementation of conjunctive queries in relational databases. In *Proc. 9th ACM Symp. on Theory of Computing*, pp. 77-90, 1977.

[4] S. Chaudhuri, M. Y. Vardi. Optimization of real conjunctive queries, In *Proc. 12th ACM Symp. on Principles of Database Systems*, pp. 59-70, 1993.

[5] P. Honeyman, R. Ladner, M. Yannakakis. Testing the universal instance assumption. *Inf. Process. Lett.*, 10(1):14-19, 1979.