Abstract:

We wish to identify the queries that can be solved with close to optimal time guarantees over relational databases. Computing all query answers requires at least linear time before the first answer (to read the input and determine the answer's existence), and then we must allow enough time to print all answers (which may be many). Thus, we aspire to achieve linear preprocessing time and constant or logarithmic time per answer. A known dichotomy classifies Conjunctive Queries into those that admit such enumeration and those that do not: the main difficulty of query answering is joining tables, which can be done efficiently if and only if the join query is acyclic. However, the join query usually does not appear in a vacuum; for example, it may be part of a larger query, or it may be applied to a database with dependencies. We show how to use this context for more efficient computation and study how the complexity changes in these settings. Next, we aspire for an even more powerful solution for query answering: a structure that simulates an array containing the query answers. Such a structure can be used for example to enumerate all answers in a statistically meaningful order or to efficiently compute a boxplot of query answers. We call this simulation random access and study for which queries random access can be achieved with near-optimal guarantees. Our results are accompanied by conditional lower bounds showing that our algorithms can be applied to all tractable queries in some cases. Among our results, we show that a union of tractable Conjunctive Queries may be intractable w.r.t. random access, but a union of intractable Conjunctive Queries may be tractable w.r.t. enumeration.