Problem Statement

Given a database schema definition $D$ and a set of queries $Q = \{Q_1, Q_2, \ldots Q_n\}$, where every $Q_i$ is annotated with a set of constraints $C_i$. Generate a database $DB$ such that: 1) $DB$ conforms $D$, and 2) the resulting cardinalities $C'_i$ of posing $Q_i$ over $DB$ approximate satisfy $C_i$.

Annotated Query

Example of symbolic query processing execution of the annotated query on the left:
1. Base table operations create the tables and populate them with unique symbols.
2. The selection operation associates constraints to the symbols in the predicate.
3. The join operation enforces reference constraints following uniform distribution.

Symbolic Query Processing

Symbolic Database Integration

The integration of two symbolic databases is performed by applying the concept of joining. The problem of integrating two tables can be modeled as the problem of finding a K-SAT-matching in a constrained bipartite graph $CBG = (U, V, E)$.

Results

In all experiments, the generated databases meet all the constraints defined in the set of input queries.

Conclusions

- The data generation process requires solving different NP-Hard problems. However, we have shown the generation of synthetic query-aware databases is possible.
- Further improvements are required to reduce the space and time complexity of our integration algorithm.

Source Code

https://github.com/mjcepedia/WAGen

References