Confidence-Aware Join Algorithms
Parag Agrawal, Jennifer Widom

Uncertain Databases
- Tuples have confidences
- Result confidence computation
- Contain input confidences
- Multiplication if independent

High confidence result tuples are more important

Threshold1
- Explore area under the stair
- IO Cost: less than 2 times optimal
- Assuming no indexes

Problem
- Leave it to the optimizer
- Treat confidence values as another column
- Result confidence computed in query

Traditional Solution
- Threshold in WHERE clause

We can do better!
- Exploit Monotonicity
  - of combining functions for join algorithms, query dependence
- Monotonicity

Join Visualization
- Nested Block Join
  - Memory size: M
  - Repeat
    - Load part of R into memory
    - Scan S and evaluate
    - Explore cross-product
  - IO Cost
    - Number of tuples read
    - Load
    - Scan

Observation
Monotonicity: C(p, q) ≥ C(r, s) if p ≥ r and q ≥ s

Experiments
- Synthetic data-sets
  - 1M, 10M tuples in each relation
  - Various conf. distributions
- Algorithms perform well
  - Not affected by confidence distributions
  - Results in paper

Top-k
- Threshold = confidence of 1st tuple
- Explore area under the stair
- Threshold value is not known

Explore in approximate order of result confidence

Future Work
- Use as operator in query plans
  - Parameter t (block size)
  - Cost estimation
  - Memory allocation
- Non-independence
  - Internal approximations
  - Monte-Carlo simulations