LMU | LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | INSTITUTE FOR INFORMATICS

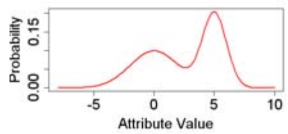
A Novel Probabilistic Pruning Approach to Speed Up Similarity Queries in Uncertain Databases

T. Bernecker, T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, A. Züfle

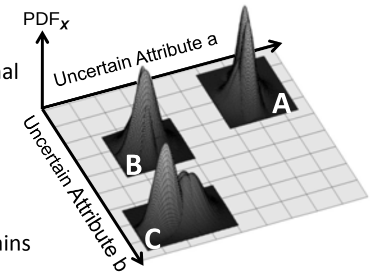
Uncertain Data Model:

- **Uncertain attribute *x***: the value of x is given by a probabilistic density function (PDF), which describes all possible values ***v*** of ***x***, associated with probability P(***x*** = ***v***)

Discrete PDF vs. **Continuous PDF**

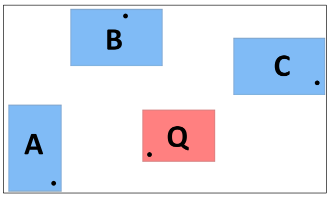


- An **uncertain object *X*** has at least d ≥ 1 uncertain attributes
- Each uncertain attribute value of ***X*** is a random variable
- ***X*** is a random variable, where the set of attribute values of ***X*** is described by a multi-dimensional probability distribution
- ***X*** has a spatial region UR_X (Uncertain Region) where
  (1) $PDF_X(\boldsymbol{t}) \geq 0$ if $\boldsymbol{t} \in UR_X$ and
  (2) $PDF_X(\boldsymbol{t}) = 0$ otherwise
- **Uncertain object database**: contains N uncertain objects

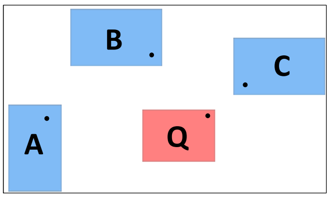


Problem Definition:

- Probabilistic k-nearest neighbor (PkNN) query: What are the k-nearest neighbors of the uncertain query object Q?



In some possible worlds, A is the nearest neighbor of Q …



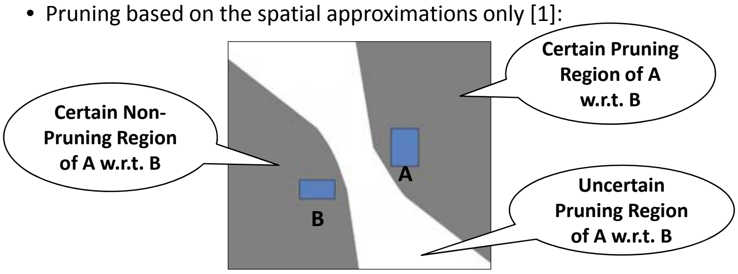
…, while in other possible worlds, that might not be the case.

General Framework:

1. **Approximation (Index):** Simplification of spatial-probabilistic keys
2. **Spatial Filter:** Filter objects according to simple spatial keys
3. **Probabilistic Filter:**
   - Derive lower/upper bounds of qualification probability (by means of simple spatial-probabilistic keys
   - Filter objects according to lower/upper probability bounds
4. **Verification:**
   - Computation of the exact probability (very expensive)
   - Monte-Carlo Sampling (in general many samples needed)

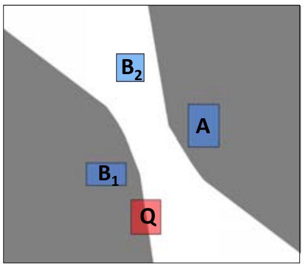
Spatial Filter:

- Pruning based on the spatial approximations only [1]:



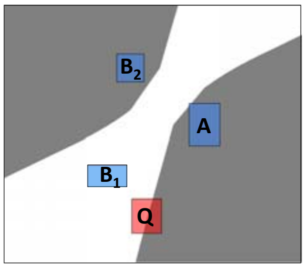
Probabilistic Filter:

- Intuition for deriving pruning probability bounds:



Lower Pruning Bound

"B_1 is closer to Q than A with a Probability of at least x %"

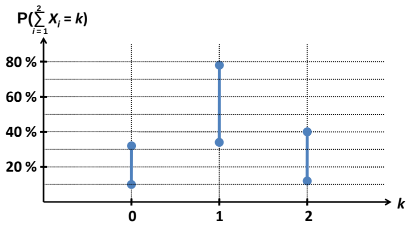


Upper Pruning Bound

"B_2 is closer to Q than A with a Probability of at most x %"

Uncertain Generating Functions:

- Assume the following pruning bounds have been derived:
  (1) B_1 prunes A with a probability of at least 0.2 and at most 0.7
  (2) B_2 prunes A with a probability of at least 0.6 and at most 0.8
- What is the probability that at least (at most, exactly) ***k*** objects prune A?
- Given the Generating Function
  (0.2x + 0.5y + 0.3) *
  (0.6x + 0.2y + 0.2)
- Expansion yields:
  $\boldsymbol{0.12x^2 + 0.34x + 0.1 +}$
  $\boldsymbol{0.22xy + 0.16y + 0.06y^2}$



- Formally let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be random predicates and let $\boldsymbol{p_i^{lb}}$ and $\boldsymbol{p_i^{ub}}$ be the respective lower and upper bounds of the probability that $\boldsymbol{X_i}$ = *true*
- Consider the expansion of the Uncertain Generating Function:

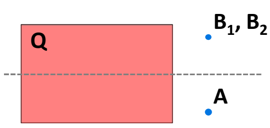
$$\prod_{i=1}^{n} p_i^{lb} x + (p_i^{ub} - p_i^{lb}) y + (1 - p_i^{ub})$$

- A lower bound of the probability that exactly ***k*** predicates are true is given by the coefficient of the term $\boldsymbol{x^k}$
- An upper bound is given by the sum of all coefficients of the term $\boldsymbol{x^m y^n}$ where $\boldsymbol{m \leq k}$ and $\boldsymbol{m + n \geq k}$

Problem of Dependencies:

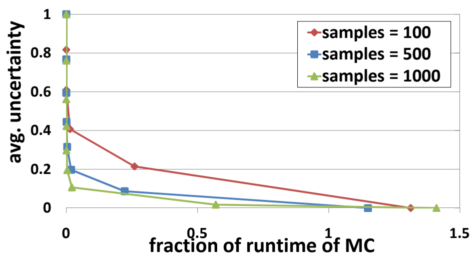
- Approximate the probability that A is the NN of Q (assuming uniform distribution in Q):

P(„B_1 is closer to Q than A“) = 0.5
P(„B_2 is closer to Q than A“) = 0.5
⇒ P(„A is NN of Q“) = 0.5 * 0.5 = 0.25 is WRONG!
⇒ P_{lb}= 0 and P_{ub}= 0.5 is CORRECT!



Evaluation:

- Comparison with Monte-Carlo-based approach:



References:

- [1] T. Emrich, H-P. Kriegel, P. Kröger, M. Renz, A. Züfle: Boosting Spatial Pruning: On Optimal Pruning of MBRs. SIGMOD Conference 2010, pp. 39-50.

INTERNATIONAL CONFERENCE ON DATA ENGINEERING ICDE 2011
HANNOVER APRIL 11-16, 2011

THE UNIVERSITY OF HONG KONG
DEPARTMENT OF COMPUTER SCIENCE