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

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

Thomas Bernecker*, Tobias Emrich*, Hans-Peter Kriegel*,
Nikos Mamoulis**, Matthias Renz* and Andreas Zuefle*

*)
Ludwig-Maximilians-Universität München (LMU)
Munich, Germany
http://www.dbs.ifi.lmu.de
{bernecker, emrich, kriegel, renz, zuefle}
@dbs.ifi.lmu.de

**)
University of Hong Kong (HKU)
Hong Kong
http://www.cs.hku.hk
nikos@cs.hku.hk

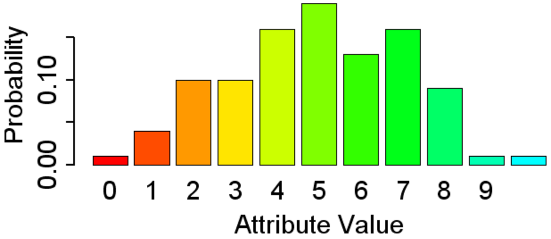
THE UNIVERSITY OF HONG KONG
DEPARTMENT OF
COMPUTER SCIENCE

Outline

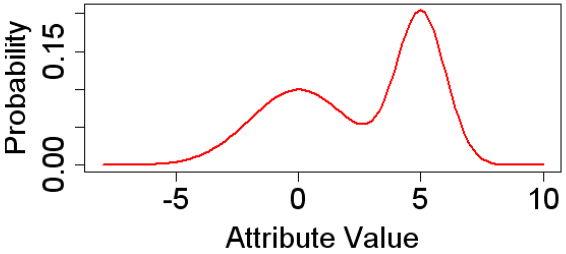
- Background
  - Uncertain Data Model
  - Similarity Queries
- Probabilistic Pruning
  - Obtaining probability bounds
  - Using probability bounds for pruning
- Evaluation

Uncertain Data Model

- Uncertain attribute
  An attribute ***x*** is uncertain if its value is given by a probabilistic density function (PDF), which describes all possible values ***v*** of ***x***, associated with probability P(***x*** = ***v***).
  - Discrete PDF (e.g. derived from missing data – See Julia's talk, derived from time series data – See Saket's talk)



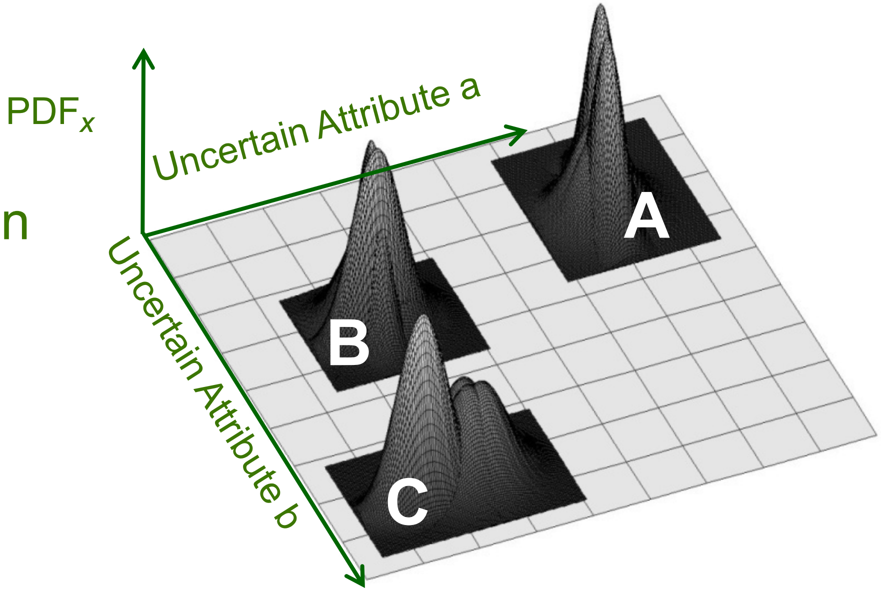
  - Continuous PDF (e.g., sensor measurement error)



Uncertain Data Model

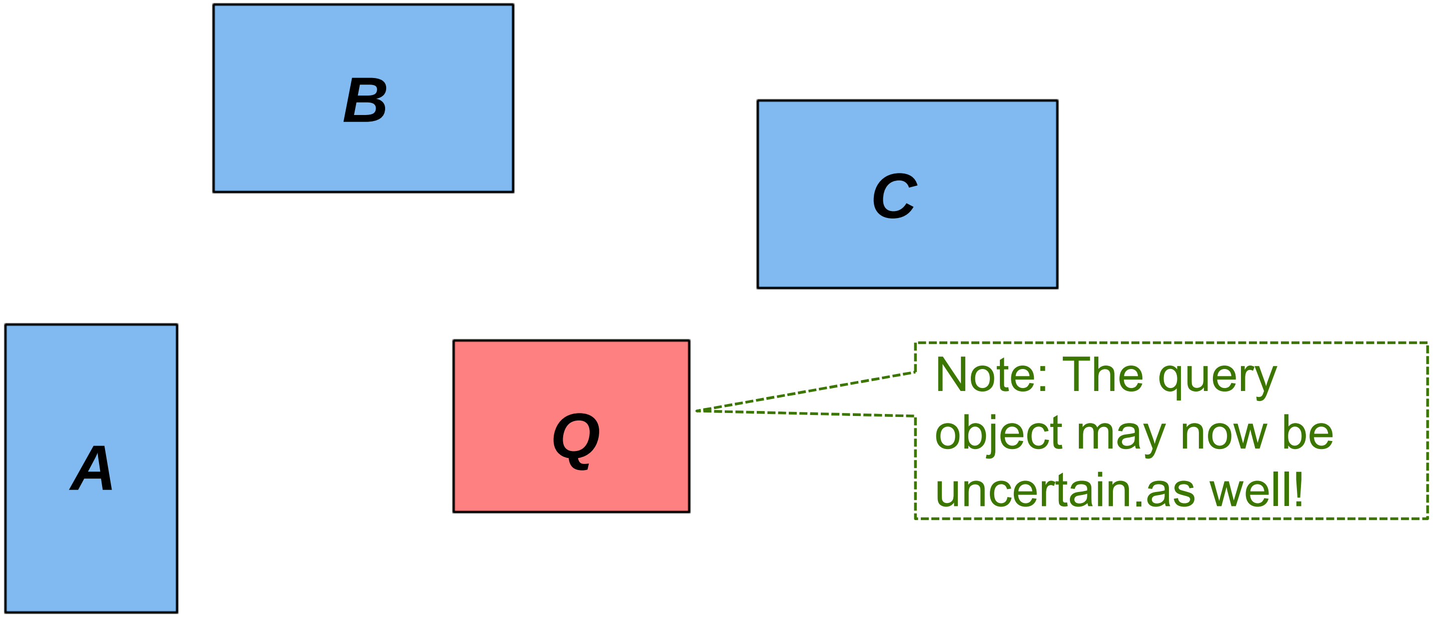
- Uncertain Object ***X***
  - Has at least d ≥ 1 uncertain attributes.
  - ***X*** is a random variable, where the set of attribute values of ***X*** is described by a multi-dimensional probability distribution.
  - ***X*** has a spatial region UR_X (Uncertain Region), where $PDF_X(t) > 0$ if $t \in UR_X$ and $PDF_X(t) = 0$ otherwise.

- Uncertain Object Database
  - Contains N uncertain objects
  - Object Independence Assumption



Probabilistic Similarity Queries

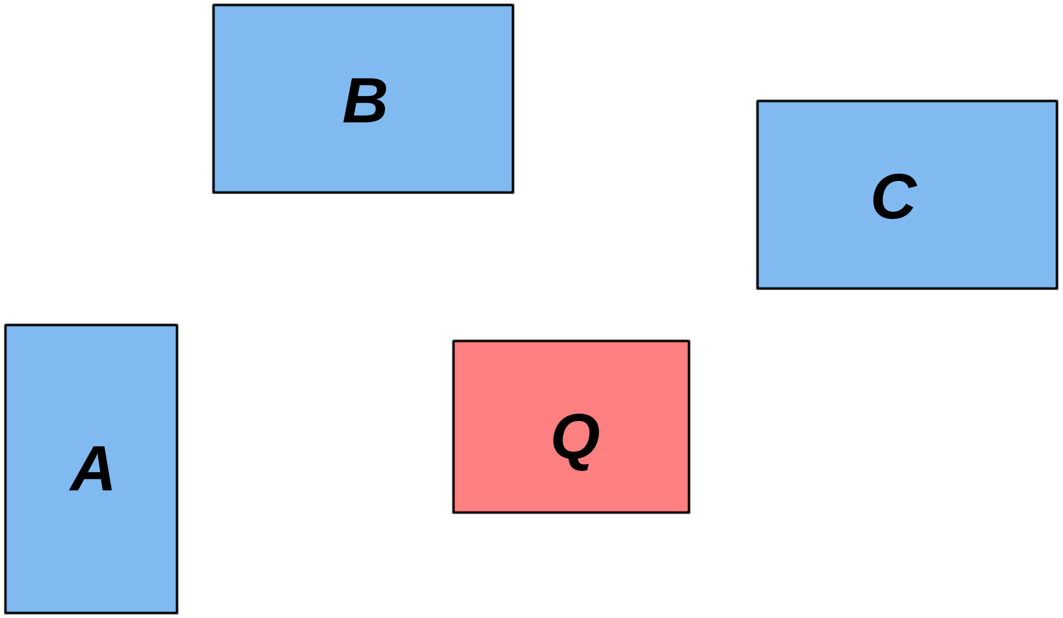
- Probabilistic k-Nearest Neighbor query
  - What are the k objects closest to Q?
- Probabilistic Similarity Ranking
  - Return all objects sorted by their distance to Q.
- Probabilistic Reverse k-Nearest Neighbor queries
- …



Note: The query object may now be uncertain.as well!

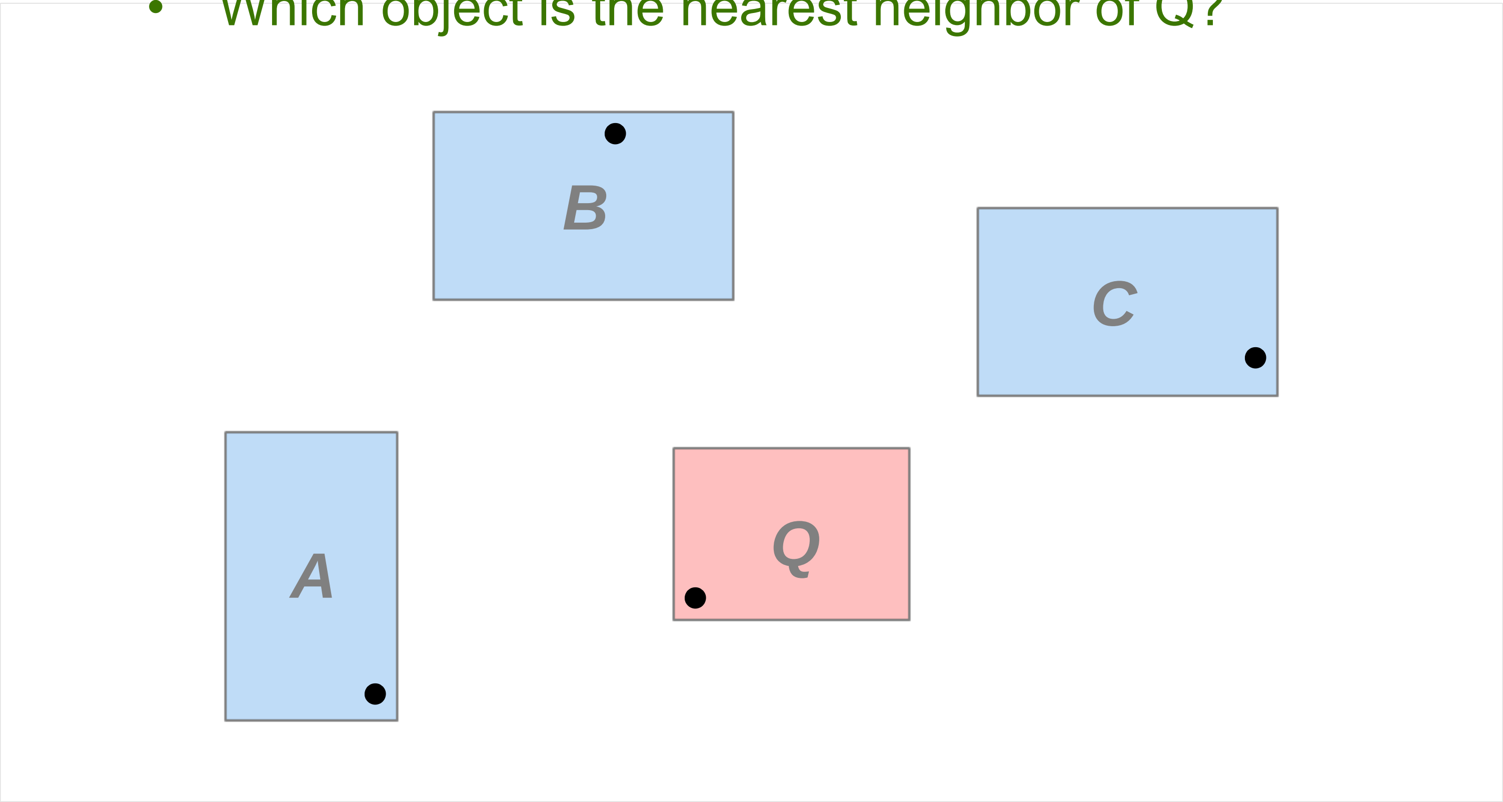
Similarity Queries: Example

- Probabilistic Nearest Neighbor query
  - Which object is the nearest neighbor of Q?



Similarity Queries: Example

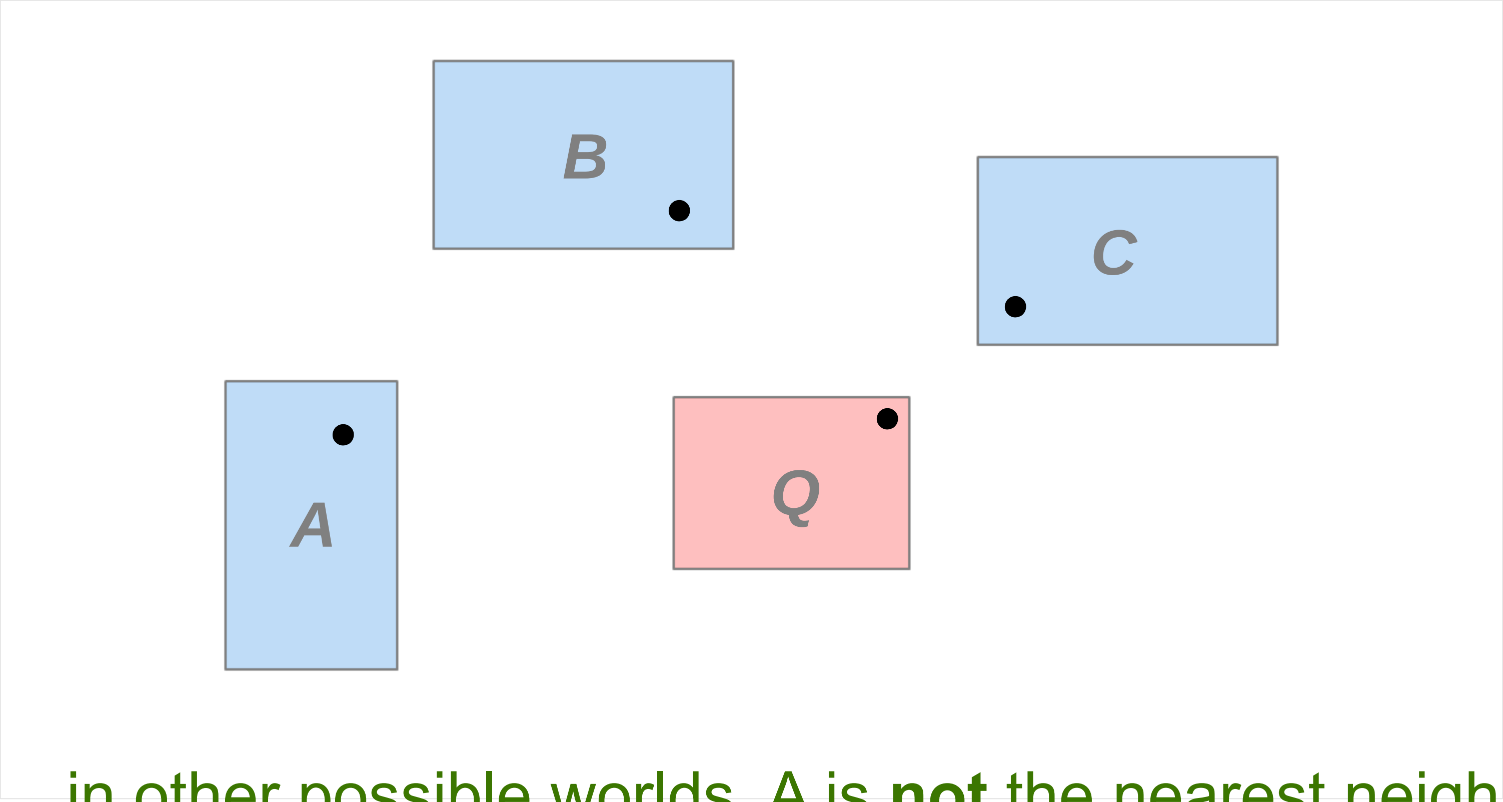
- Probabilistic Nearest Neighbor queries
  - Which object is the nearest neighbor of Q?



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

Similarity Queries: Example

- Probabilistic Nearest Neighbor queries
  - Which object is the nearest neighbor of Q?



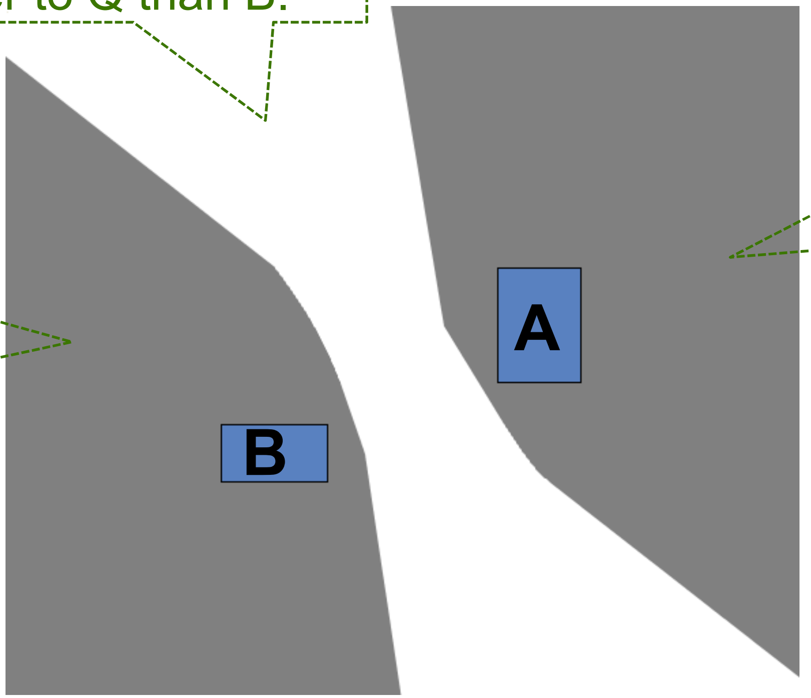
…in other possible worlds, A is **not** the nearest neighbor of Q.

General Framework

- Efficient probabilistic similarity search:
  - Approximation (Index)
    - Simplification of spatial-probabilistic keys
  - Spatial Filter
    - Filter objects according to simple spatial keys
  - Probabilistic Filter
    - Derive lower/upper bounds of qualification probability (by means of simple spatial-probabilistic keys)
    - Filter objects according to lower/upper probability bounds
  - Verification
    - Computation of the exact probability (very expensive)
    - Monte-Carlo Sampling (many samples required)

Spatial Filter

Pruning based on rectangular approximations only [1].



For any Q in this region, A may possibly be closer to Q than B.

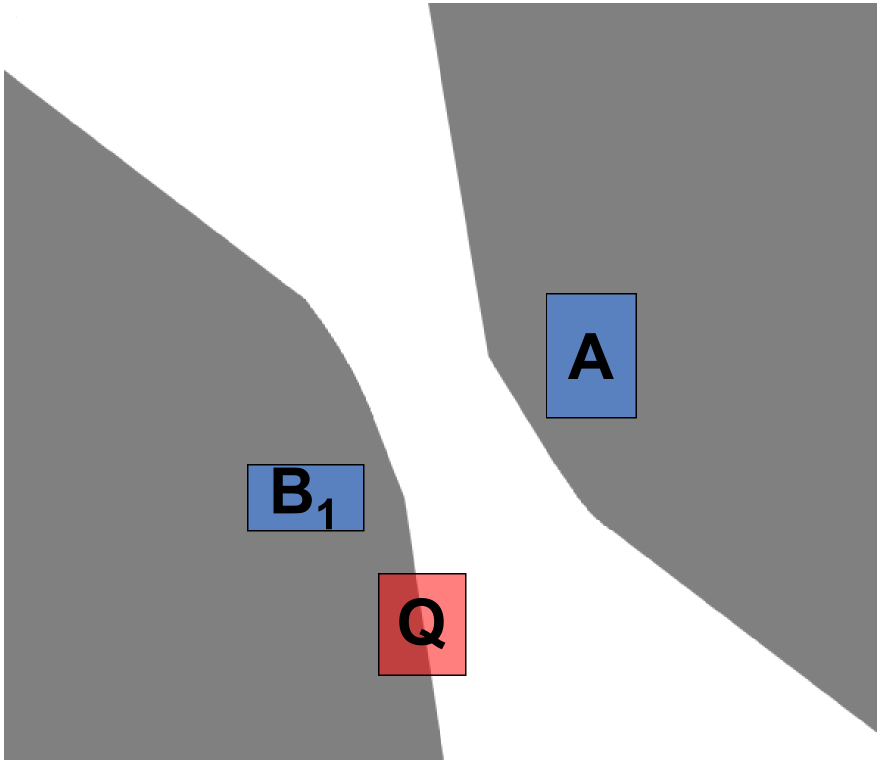
For any Q in this region, A is closer to Q than B.

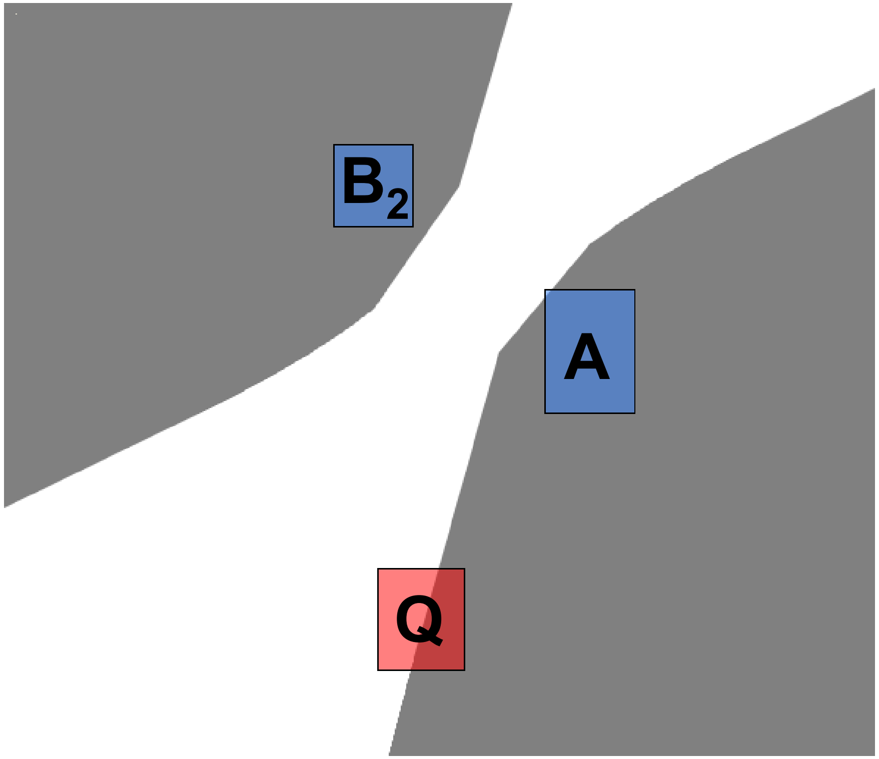
For any Q in this region, A is not closer to Q than B.

[1] Tobias Emrich, Hans-Peter Kriegel, Peer Kröger, Matthias Renz, Andreas Züfle: Boosting Spatial Pruning: On Optimal Pruning of MBRs. SIGMOD Conference 2010: 39-50

Probabilistic Pruning

How many objects are closer to Q than A?





Lower Probability Bound

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

Upper Probability Bound

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

Uncertain Generating Functions

- What we have now is:
  - ***B_1*** is closer to **Q** than **A** with a probability of at least ***p_1^{lb}*** and at most ***p_1^{ub}***
  - ***B_2*** is closer to **Q** than **A** with a probability of at least ***p_2^{lb}*** and at most ***p_2^{ub}***
  - ...

- How can we derive the probability that at least (at most, exactly) k objects are closer to Q than A?

Uncertain Generating Functions

- Let **φ** be a predicate and let X_1, …, X_n be uncertain objects. Let p_i^{lb} and p_i^{ub} be lower and upper bounds of the probability that X_i satisfies **φ**.

- How many objects satisfy **φ**?

- We consider the following generating function:

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

Example

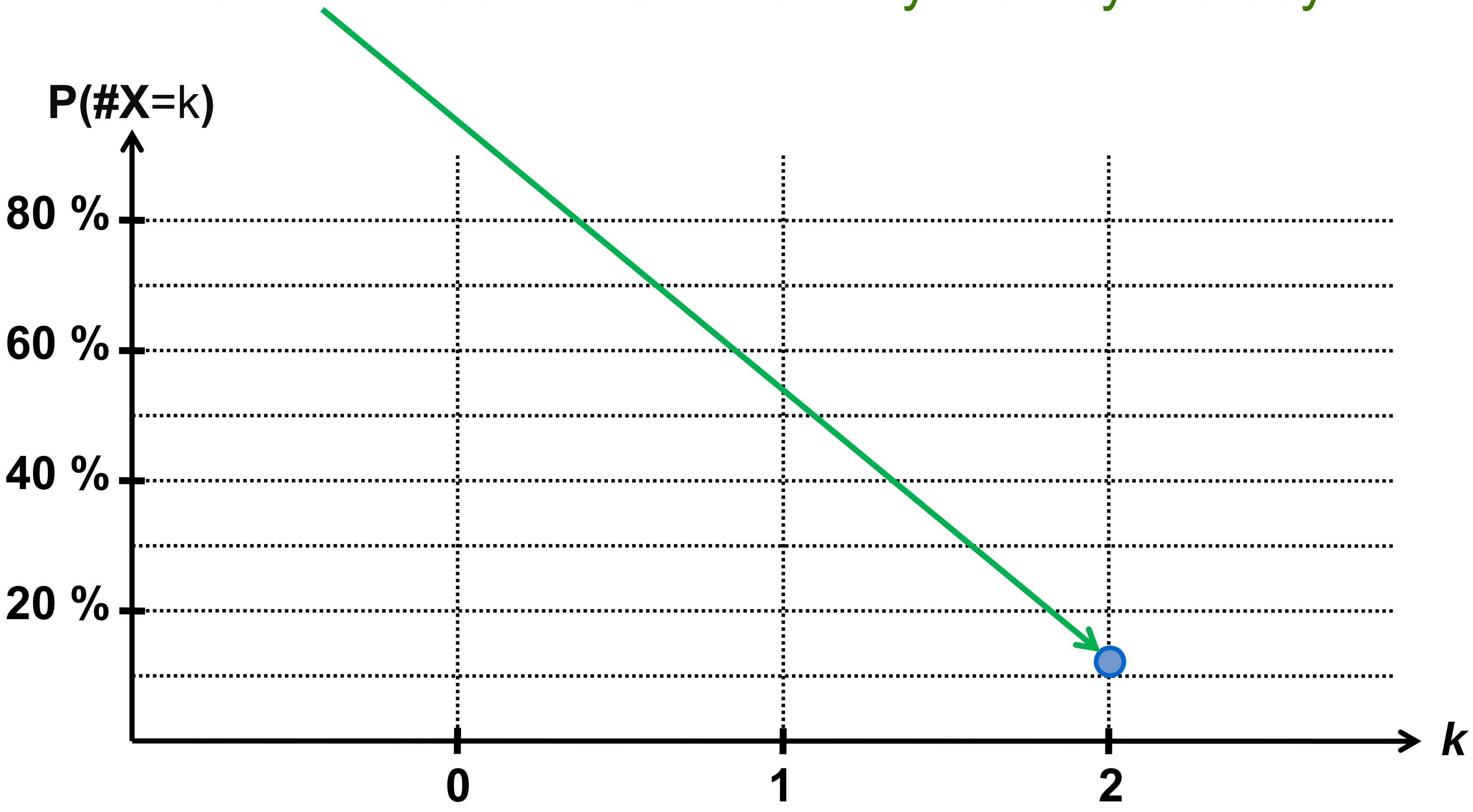
- Assume the following probability bounds have been derived:
  - X_1 satisfies **φ** with a probability of at least 0.2 and at most 0.5
  - X_2 satisfies **φ** with a probability of at least 0.6 and at most 0.8

- What is the probability that the number **#X** of objects that satisfy **φ** is at least (at most, exactly) **k**?
  - Consider the following Generating Function:
    $(0.2x + 0.3y + 0.5) * (0.6x + 0.2y + 0.2)$
  - Expansion yields:
    $0.12x^2 + 0.34x + 0.1 + 0.22xy + 0.16y + 0.06y^2$

Uncertain Generating Functions

- Expansion yields:

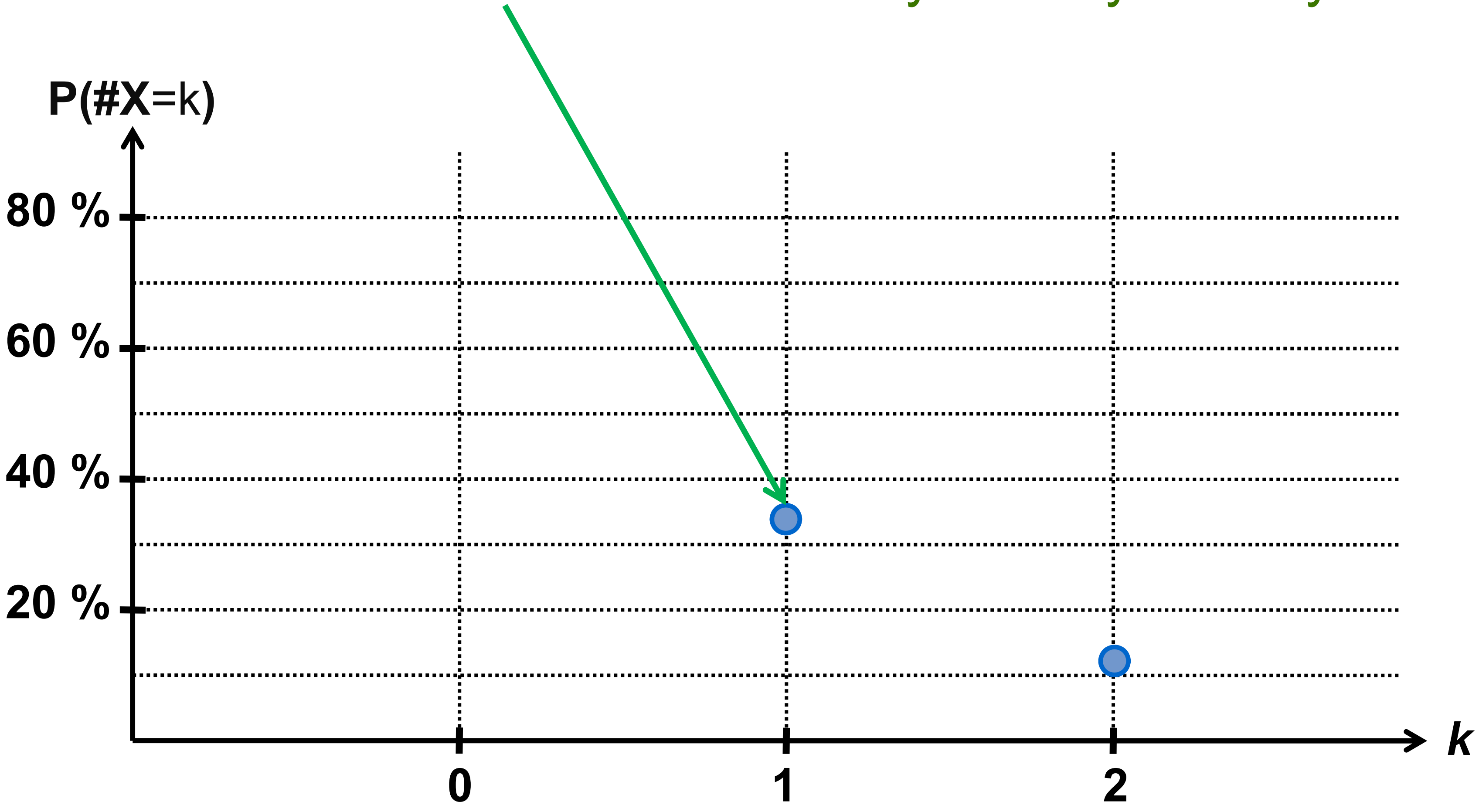
  $\mathbf{0.12}x^2 + 0.34x + 0.1 + 0.22xy + 0.16y + 0.06y^2$



Uncertain Generating Functions

– Expansion yields:

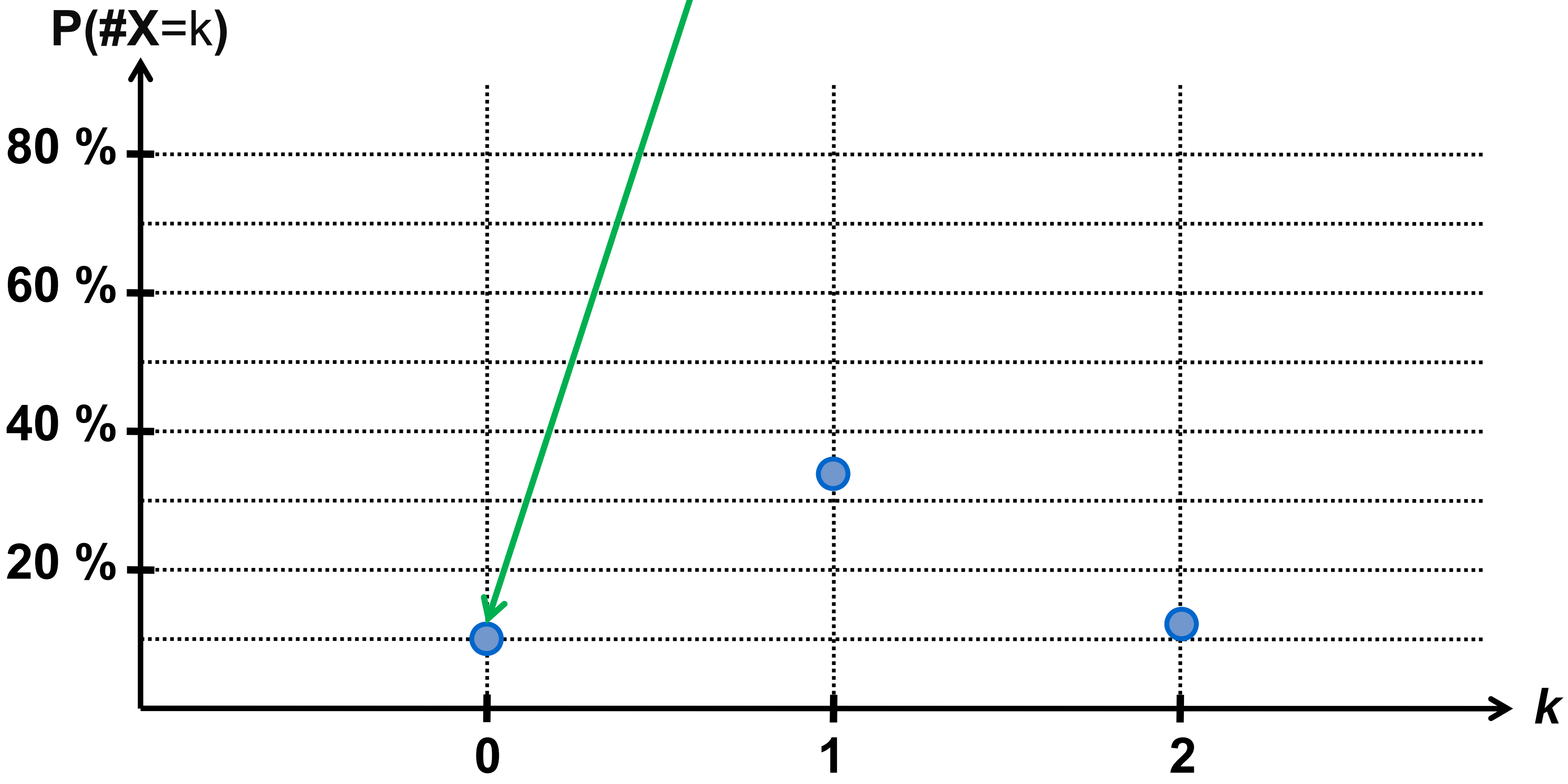
$0.12x^2 + \mathbf{0.34}x + 0.1 + 0.22xy + 0.16y + 0.06y^2$



Uncertain Generating Functions

– Expansion yields:

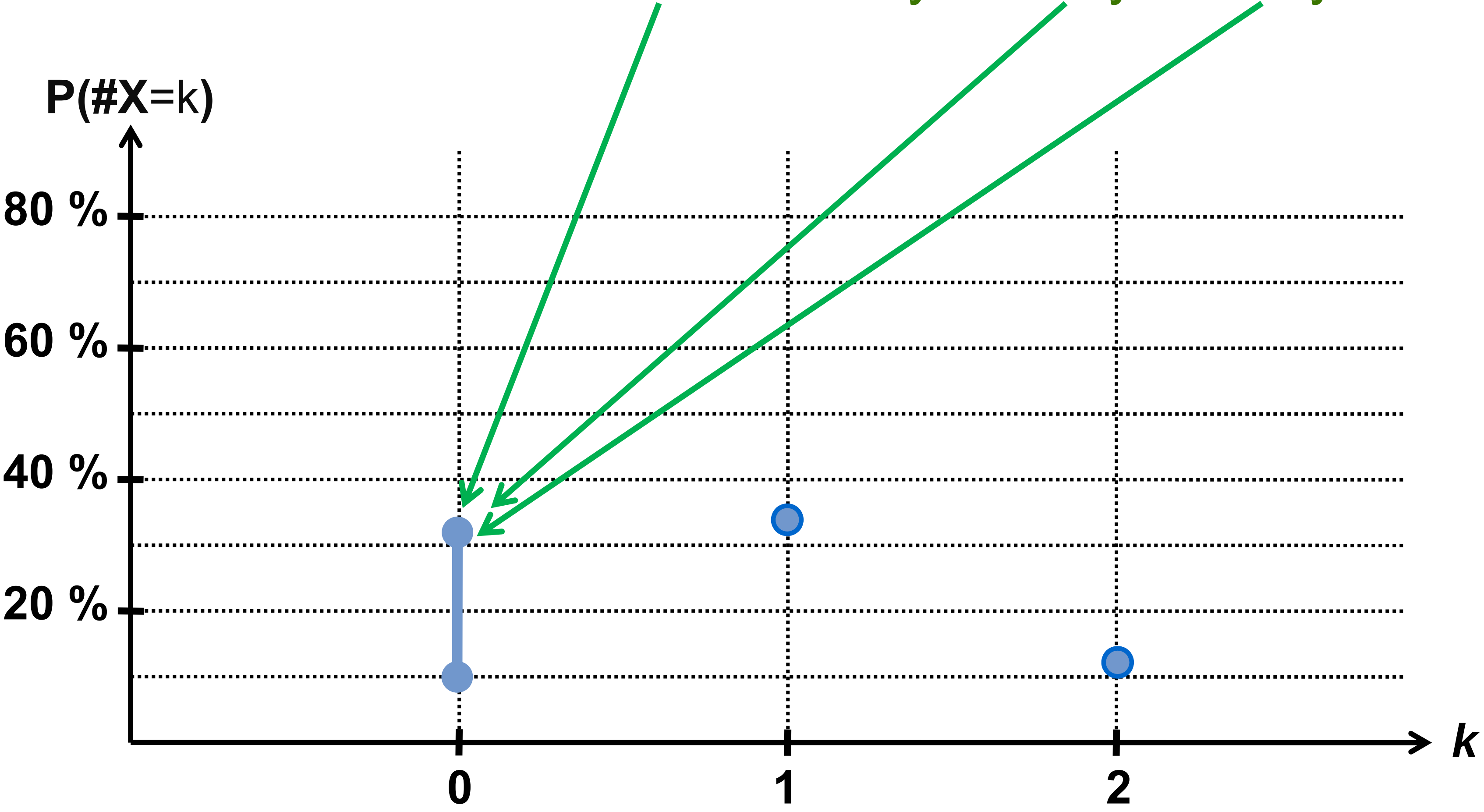
$0.12x^2 + 0.34x + \mathbf{0.1} + 0.22xy + 0.16y + 0.06y^2$



Uncertain Generating Functions

- Expansion yields:

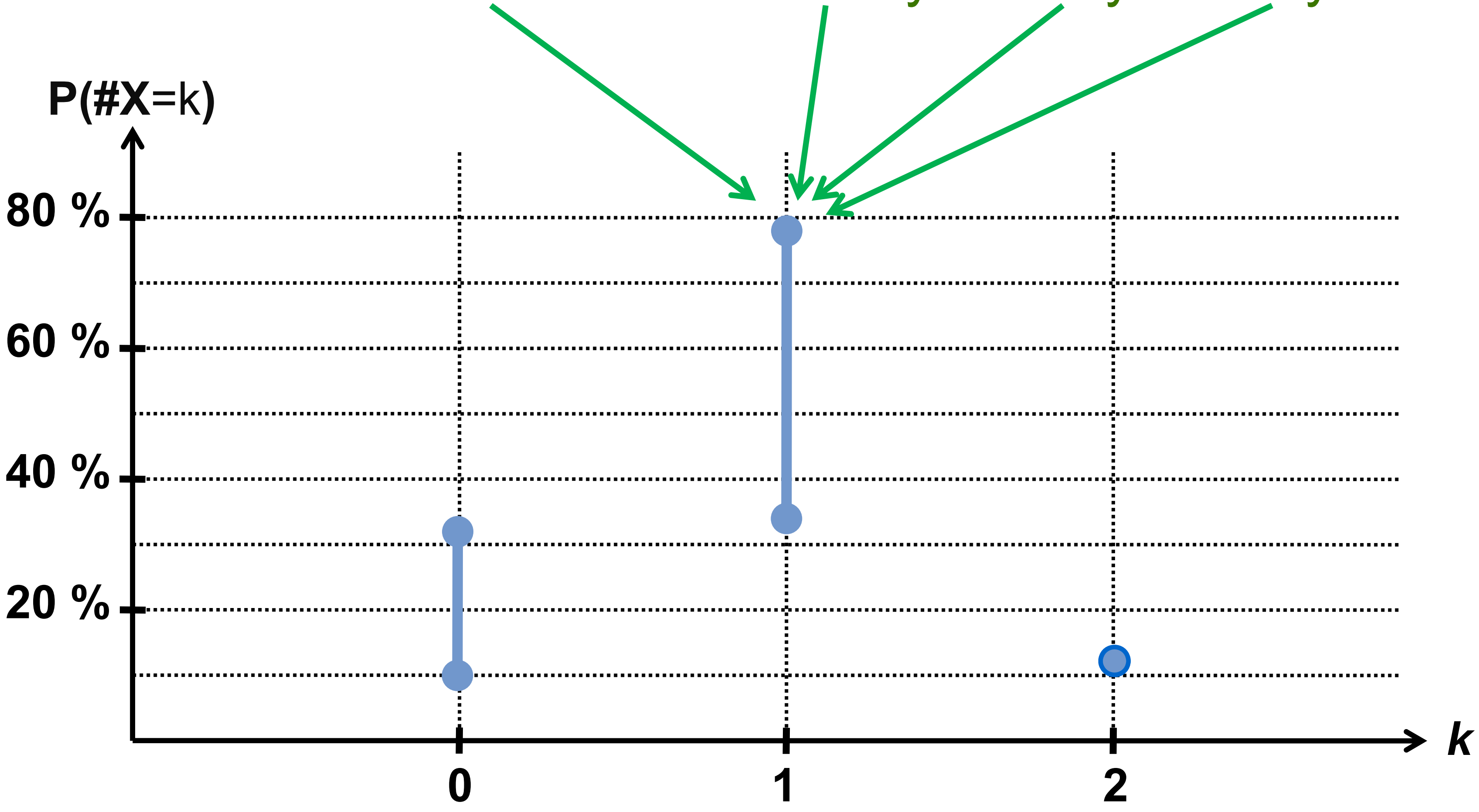
  $0.12x^2 + 0.34x + \mathbf{0.1} + 0.22xy + \mathbf{0.16}y + \mathbf{0.06}y^2$



Uncertain Generating Functions

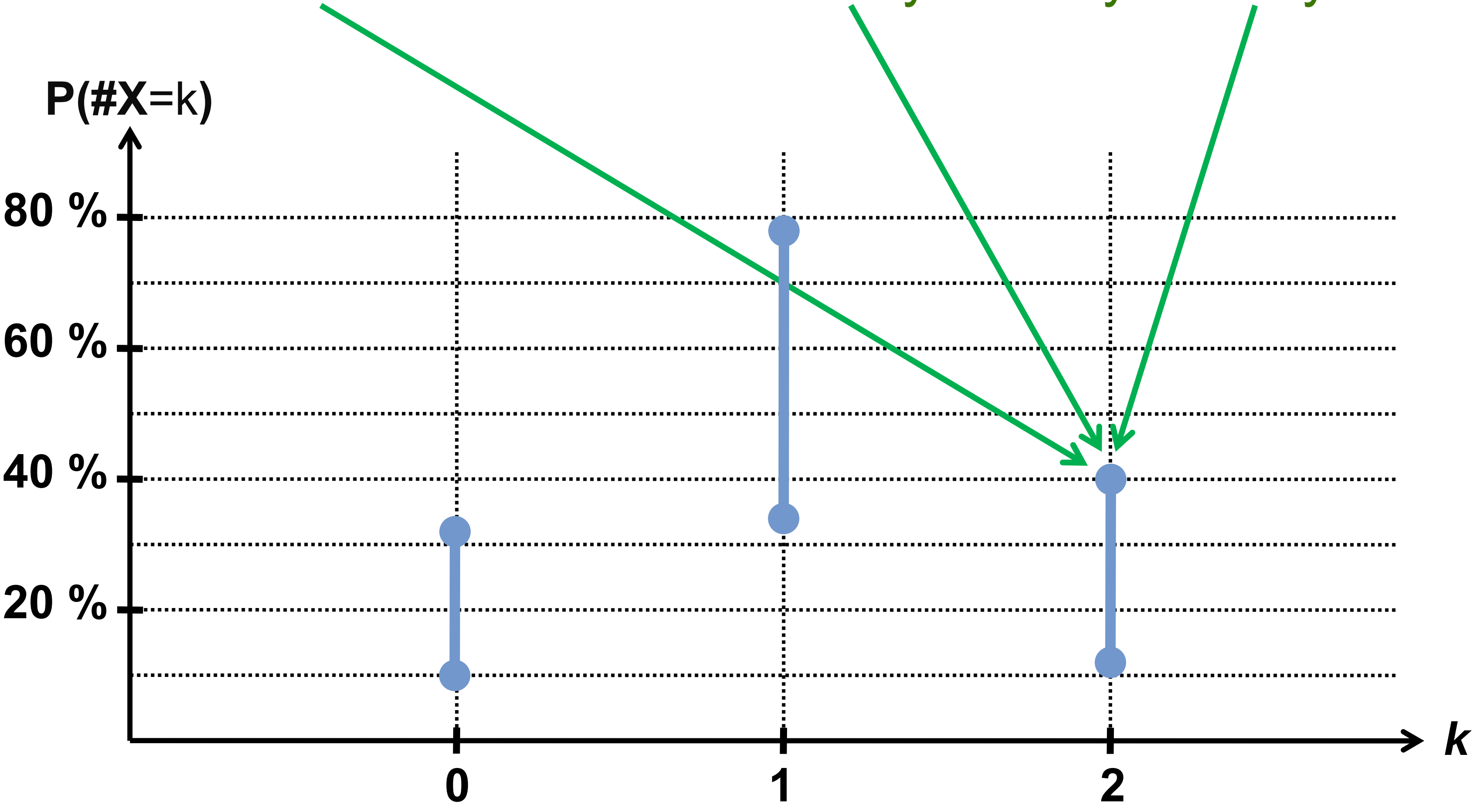
– Expansion yields:

$0.12x^2 + \mathbf{0.34}x + 0.1 + \mathbf{0.22}xy + \mathbf{0.16}y + \mathbf{0.06}y^2$



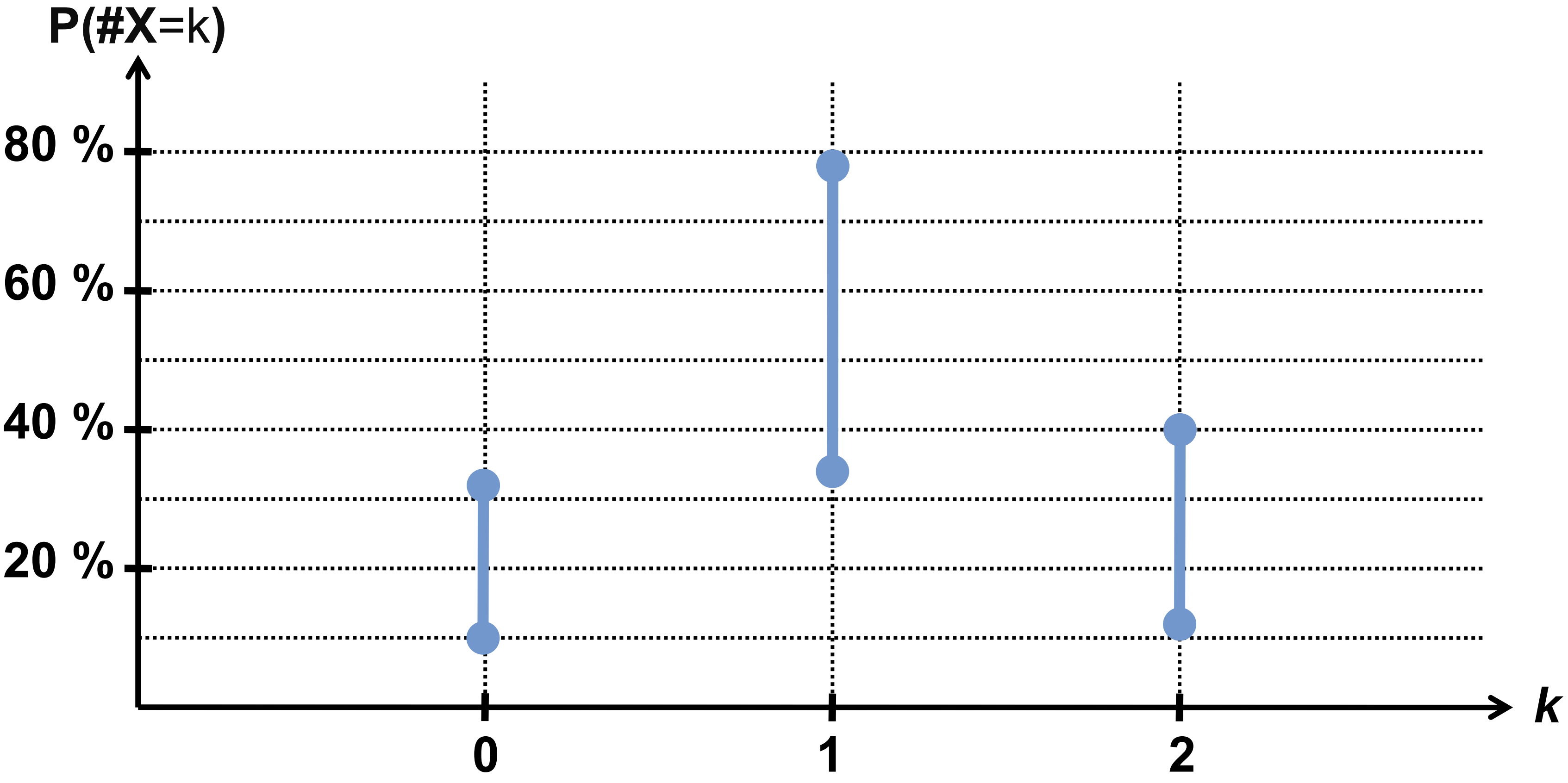
Uncertain Generating Functions

- Expansion yields:
  $\mathbf{0.12}x^2 + 0.34x + 0.1 + \mathbf{0.22}xy + 0.16y + \mathbf{0.06}y^2$

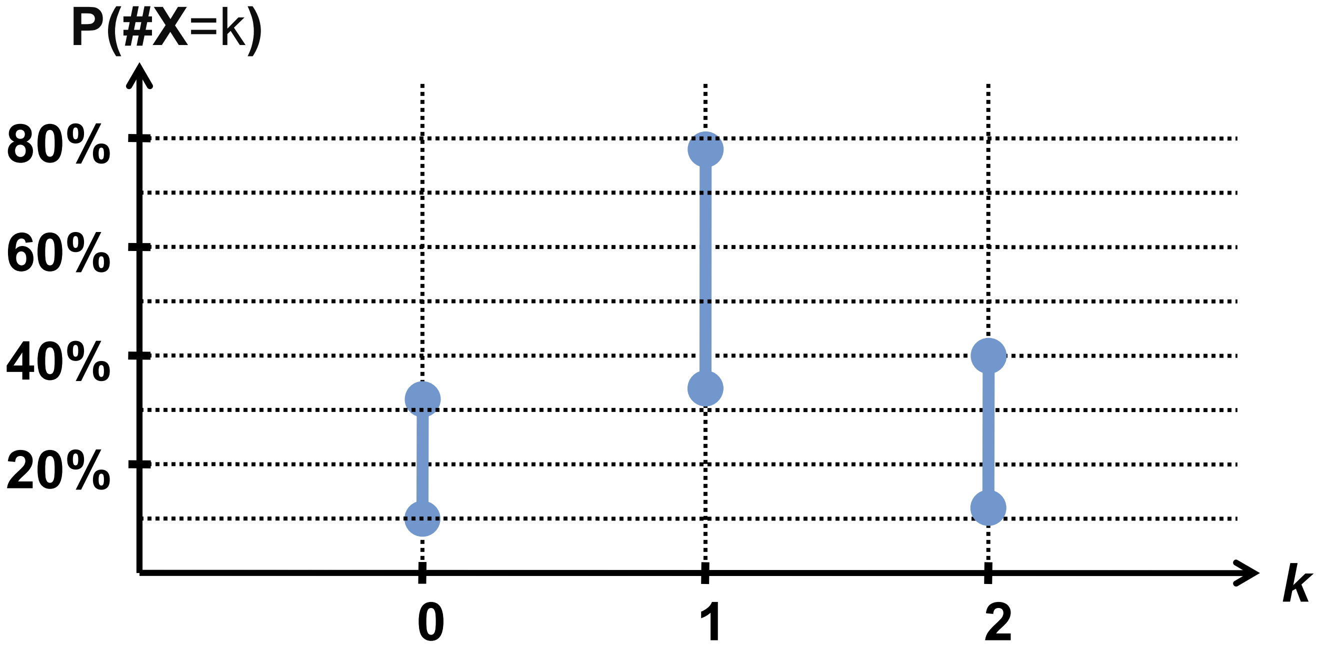


Approximated PDF

The result is an approximated PDF of **#X**.



Uncertain Generating Functions



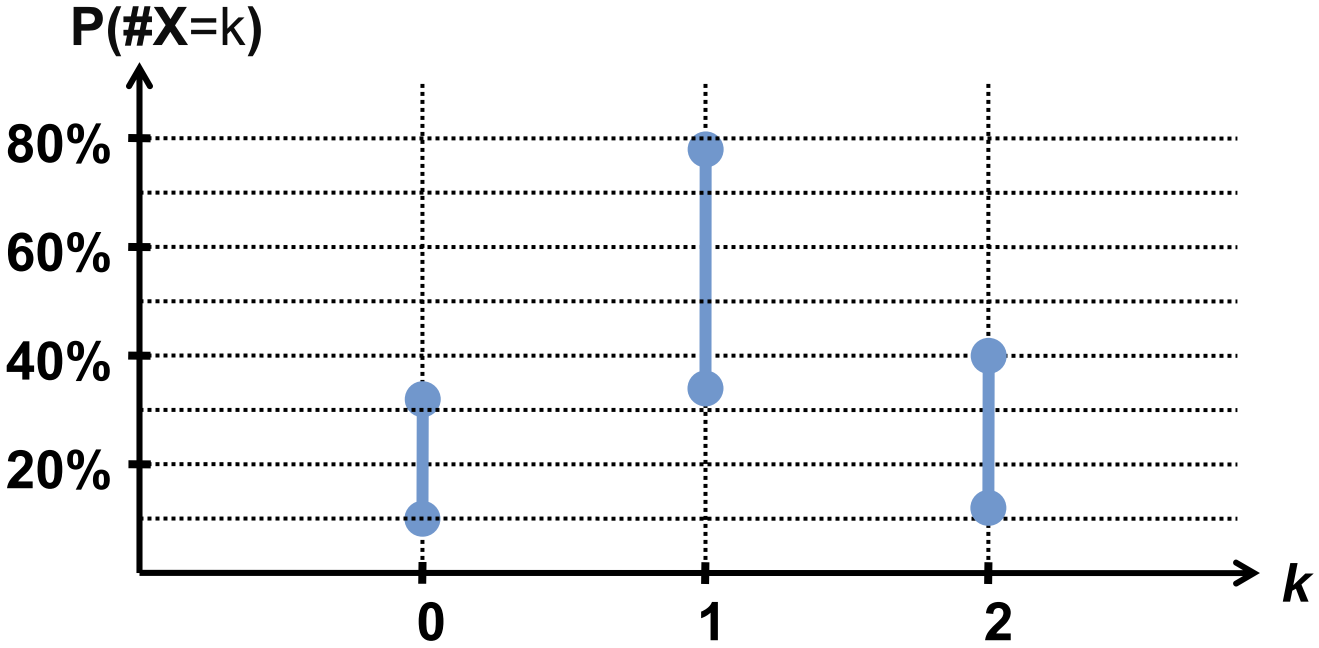
Now let **#X** denote the number of objects that are closer to **Q** than **A**.

The pdf of **#X** corresponds directly of the similarity rank of **A** to **Q**.

Example Query: Return all objects that are the nearest neighbor of **Q** with a probability of at least 50%.

- **A** can be pruned.

Uncertain Generating Functions



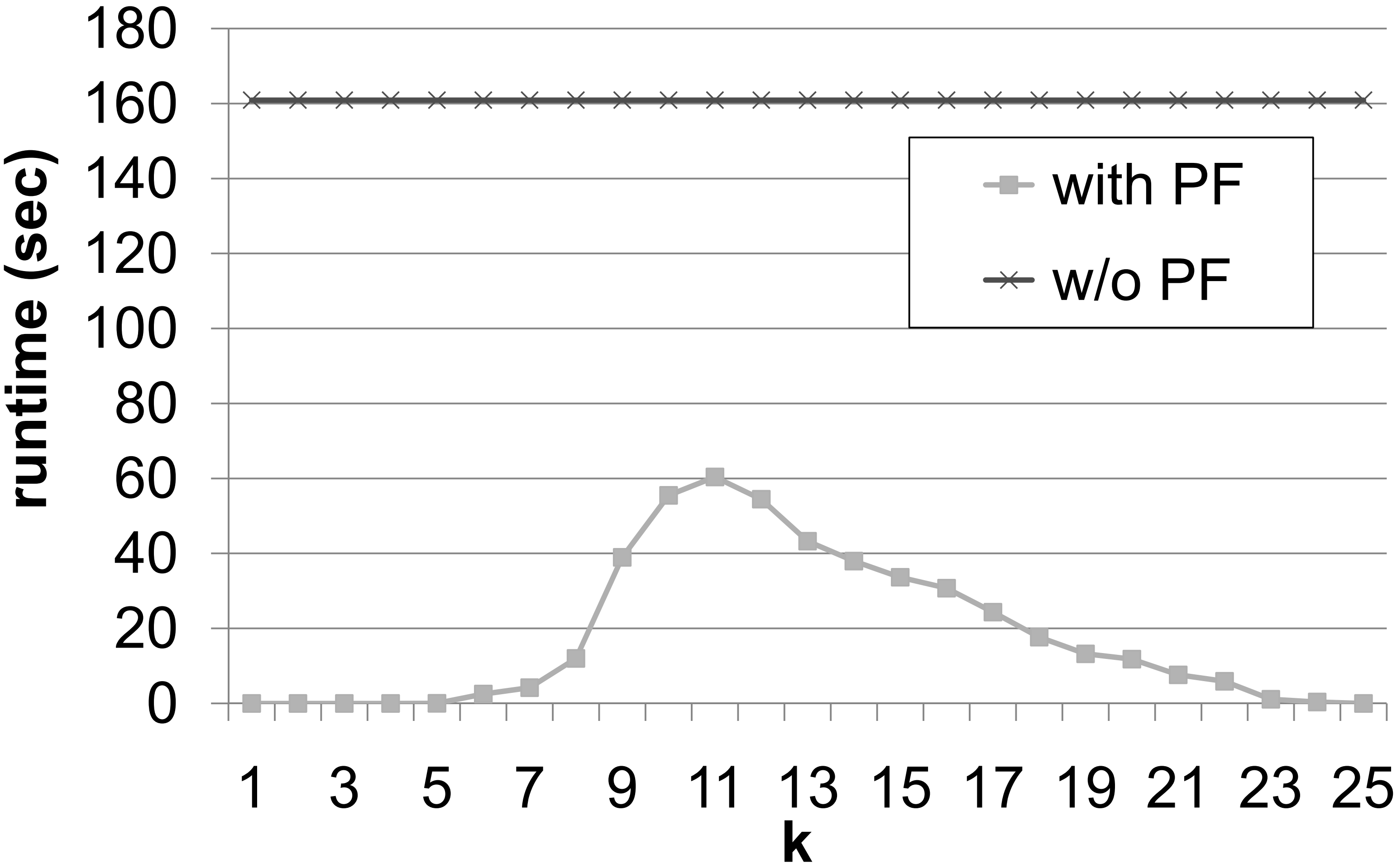
Now let **#X** denote the number of objects that are closer to **Q** than **A**.

The pdf of **#X** corresponds directly of the similarity rank of **A** to **Q**.

Example Query: Return the most likely rank of each object.

- For **A**, Rank 1 can be pruned.

Evaluation


180
160
140
120
100
80
60
40
20
0
runtime (sec)
1
3
5
7
9
11
13
15
17
19
21
23
25
k
with PF
w/o PF

Summary

- Algorithm to handle probabilistic similarity queries with an uncertain query object
- Use of spatial pruning technique to obtain probability bounds
- Efficient and correct accumulation of bounds using uncertain generation functions