Probabilistic Skylines on Uncertain Data

BIN JIANG

UNIVERSITY OF NEW SOUTH WALES

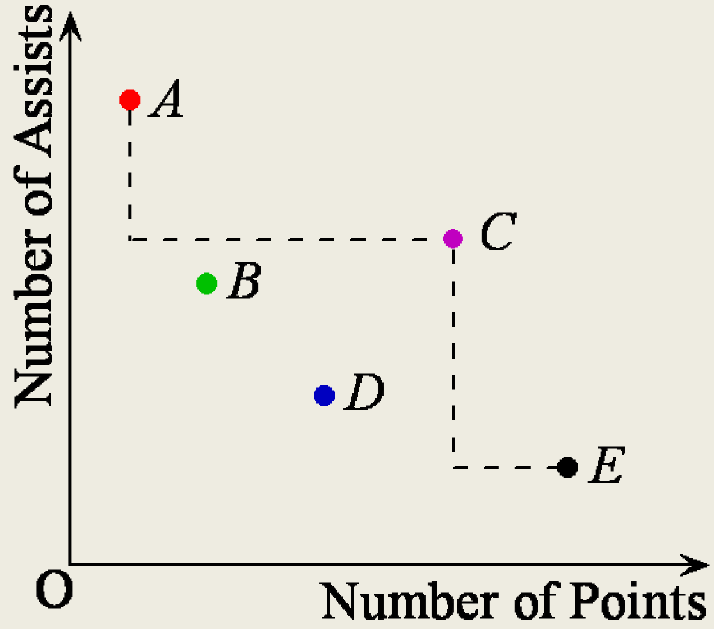
Collaborators:
Jian Pei (SFU)
Xuemin Lin (UNSW & NICTA)
Yidong Yuan (UNSW & NICTA)

Outline

- Skyline Analysis on Uncertain Data
- Related Work
- Probabilistic Skyline Model
- Probabilistic Skyline Computation
- Experiments
- Conclusions

Conventional Skylines

- n-dimensional numeric space $D = (D_1, \ldots, D_n)$
- Large values are preferable
- Two points, u dominates v ($u \succ v$), if
  - $\forall D_i\ (1 \le i \le n),\ u.D_i \ge v.D_i$
  - $\exists D_j\ (1 \le j \le n),\ u.D_j > v.D_j$
- Given a set of points S, skyline = $\{u \mid u \in S$ and u is not dominated by any other point$\}$
- Example
  - $C \succ B$, $C \succ D$
  - skyline = $\{A, C, E\}$

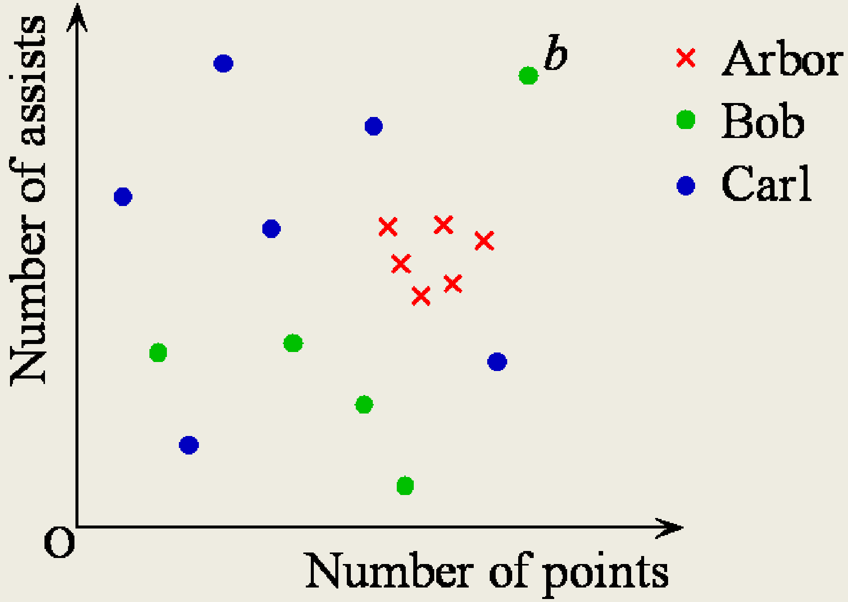


Related Work – Skyline

- Skyline computation:
  - Non-index: BNL [ICDE'01], DC [ICDE'01], SFS [ICDE'03], LESS [VLDB'05], …
  - Index: Bitmap [VLDB'01], Index [VLDB'01], NN [VLDB'02], BBS [SIGMOD'03], …
- Skyline variants:
  - Skyline cubes [VLDB'05, SIGMOD'06, ICDE'07]
  - Subspace skyline: SUBSKY [ICDE'06]
  - …

Skylines on Uncertain Data

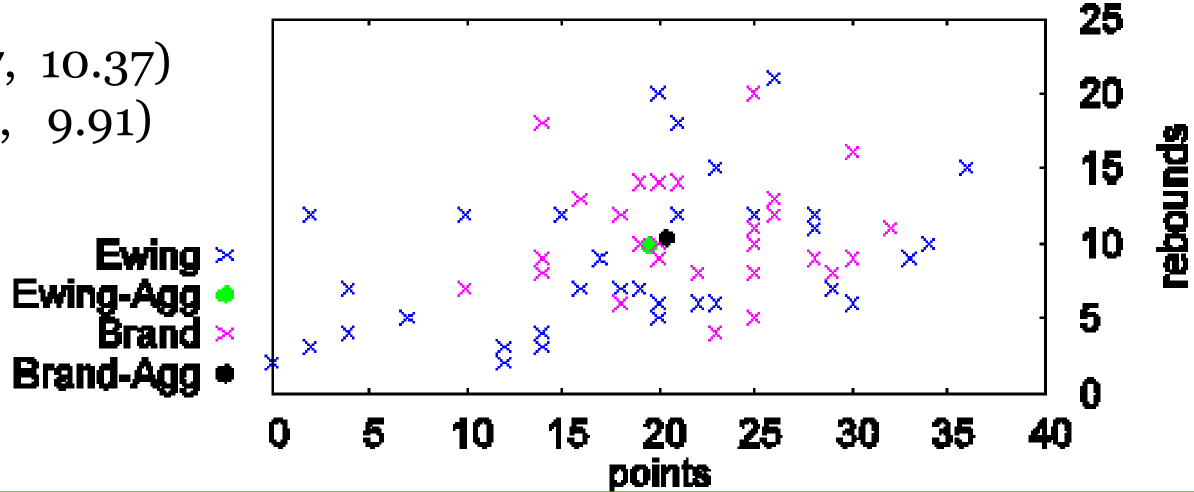
- Consider game-by-game statistics
- Conventional methods compute the skyline on
  - Separate game records
  - Aggregate: mean or median
- Limitations
  - Biased by outliers
  - Lose data distributions
- Probabilistic skylines
  - An instance has a probability to represent the object
  - An object has a probability to be in the skyline



339,721 game records of 1,313 players in 3d-space: (points, assists, rebounds)
red color : the conventional skyline computed on the aggregate statistics

Player Name	Skyline Probability	Player Name	Skyline Probability	Player Name	Skyline Probability
LeBron James	0.350699	Dwyane Wade	0.199065	Steve Francis	0.131061
Dennis Rodman	0.327592	Tracy Mcgrady	0.198185	Dirk Nowitzki	0.130301
Shaquille O'Neal	0.323401	Grant Hill	0.191164	Paul Pierce	0.127079
Charles Barkley	0.309311	John Stockton	0.183591	Gary Payton	0.126328
Kevin Garnett	0.302531	David Robinson	0.177437	Baron Davis	0.125298
Jason Kidd	0.293569	Stephon Marbury	0.16683	Vince Carter	0.122946
Allen Iverson	0.269871	Tim Hardaway	0.166206	Antoine Walker	0.121745
Michael Jordan	0.250633	Magic Johnson	0.151813	Steve Nash	0.115874
Tim Duncan	0.241252	Chris Paul	0.149264	Andre Miller	0.11275
Karl Malone	0.239737	Gilbert Arenas	0.142883	Isiah Thomas	0.11076
Chris Webber	0.22153	Clyde Drexler	0.138993	Elton Brand	0.10966
Kevin Johnson	0.208991	Patrick Ewing	0.13577	Scottie Pippen	0.108941
Hakeem Olajuwon	0.203641	Rod Strickland	0.135735	Dominique Wilkins	0.104323
Kobe Bryant	0.200272	Brad Daugherty	0.133572	Lamar Odom	0.101803

Brand-Agg (20.39, 2.67, 10.37)
Ewing-Agg (19.48, 1.71, 9.91)

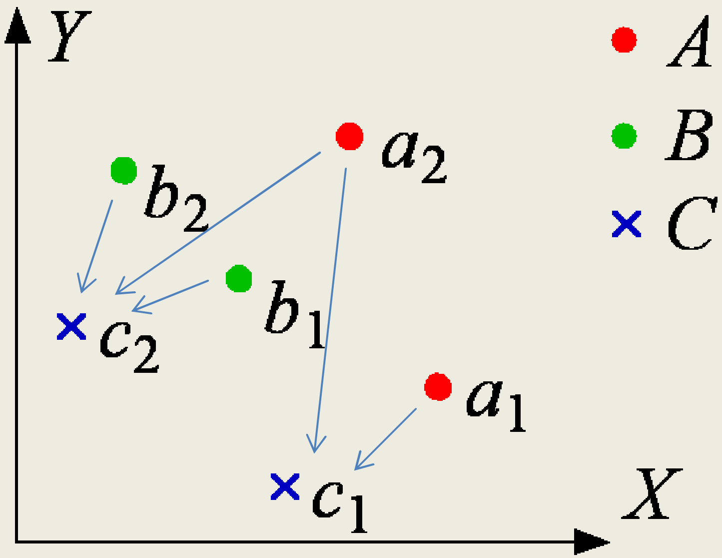


Related Work – Uncertain Data

- Uncertain Data
  - Survey [PODS'07]
  - Probabilistic range query [VLDB'04]
  - U-Tree [VLDB'05]
  - Probabilistic similarity join [DASFAA'06]
  - …
- An uncertain object is represented as
  - Continuous case: a probabilistic density function (PDF)
  - Discrete case: a set of instances, each takes a probability to appear
    - $U = \{u_1, \ldots, u_n\}$, $0 < p(u_i) \leq 1$ and $\sum_{1 \leq i \leq n} p(u_i) = 1$
    - Without loss of generality, assume equal probability, $p(u_i) = 1 / |U|$

Probabilistic Skyline Model

- Example
  - A set of object $S = \{A, B, C\}$
  - Each instance takes equal probability (0.5) to appear



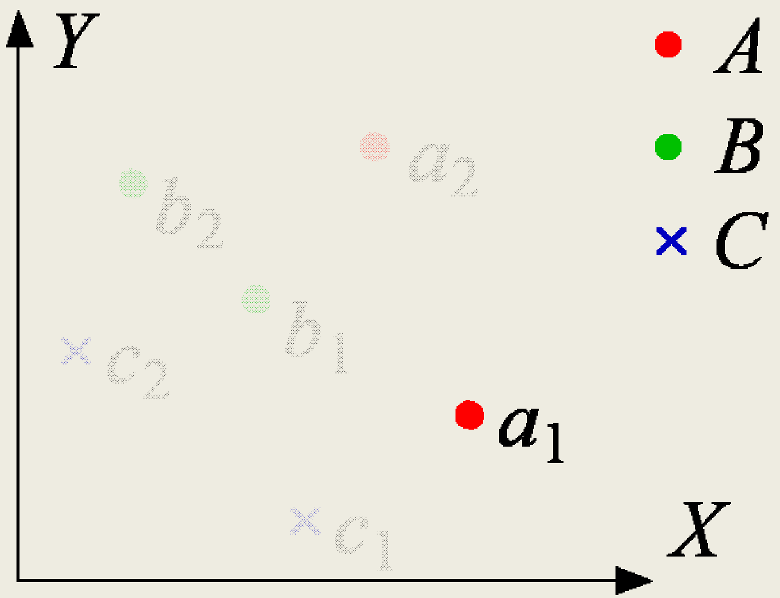
- Probabilistic Dominance
  - $Pr(A \succ C) = 3/4$
  - $Pr(B \succ C) = 1/2$
  - $Pr((A \succ C) \vee (B \succ C)) = 1$
  - $Pr(C \text{ is in the skyline}) \neq (1 - Pr(A \succ C)) \times (1 - Pr(B \succ C))$
  - Probabilistic dominance $\not\Longrightarrow$ Probabilistic skyline

Skyline Probability

- Possible world: $W = \{a_i, b_j, c_k\}$ $(i, j, k = 1 \text{ or } 2)$
- $Pr(W) = 0.5 \times 0.5 \times 0.5 = 0.125$
- $\sum_{W \in \Omega} Pr(W) = 1$

- $SKY(\{a_1, b_1, c_1\}) = \{a_1, b_1\}$
- A and B are in $SKY(\{a_1, b_1, c_1\})$

- B is in the skyline of $\{a_1, b_1, c_1\}$, $\{a_1, b_1, c_2\}$, $\{a_1, b_2, c_1\}$, and $\{a_1, b_2, c_2\}$
- $Pr(B) = 4 \times 0.125 = 0.5$
- $Pr(A) = 1$, $Pr(C) = 0$



Problem Statement

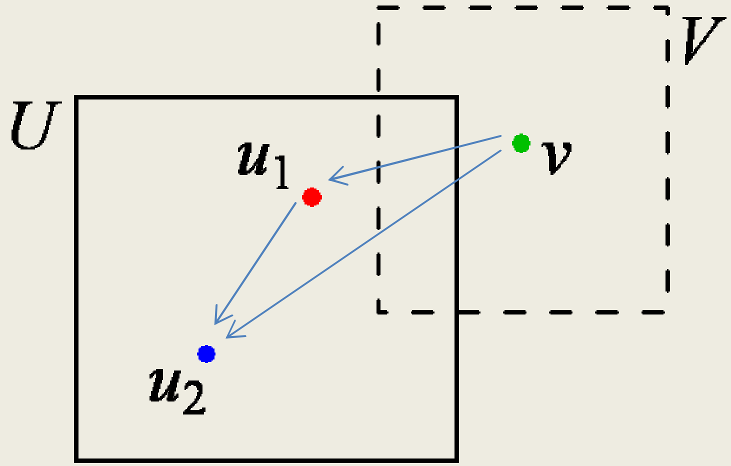
- Skyline probability: $Pr(U) = \sum_{U \in SKY(W)} Pr(W)$
- For object: $Pr(U) = \frac{1}{|U|} \sum_{u \in U} \prod_{V \neq U} (1 - \frac{|\{v \in V \mid v \phi U\}|}{|V|})$
- For instance: $Pr(u) = \prod_{V \neq U} (1 - \frac{|\{v \in V \mid v \phi u\}|}{|V|})$
- $Pr(U) = \frac{1}{|U|} \sum_{u \in U} Pr(u)$
- *p*-skyline = $\{U \mid Pr(U) \geq p\}$ for a given threshold p

Probabilistic Skyline Computation

- Iteration: Bounding-Pruning-Refining
- Bounding
  - Bound $Pr(u)$: lower bound $Pr^{-}(u)$ and upper bound $Pr^{+}(u)$
  - Bound $Pr(U)$: $Pr(U) = \frac{1}{|U|}\sum_{u \in U} Pr(u)$
- Pruning
  - In p-skyline if lower bound $Pr^{-}(U) \geq p$
  - Not in p-skyline if upper bound $Pr^{+}(U) < p$
- Refining
  - Bottom-up method
  - Top-down method

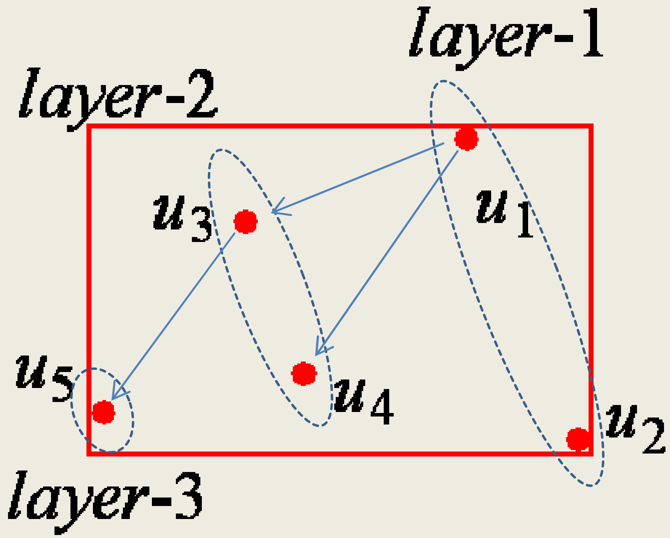
Bottom-Up Method

- Key idea: sort the instances of an object according to dominance relation, s.t.,
  their skyline probabilities are in descending order
- Dominance ➔ partial order of skyline probabilities
- Lemma
  - Two instances u_1 and $u_2 \in U$, if $u_1 \succ u_2$, then $Pr(u_1) \geq Pr(u_2)$



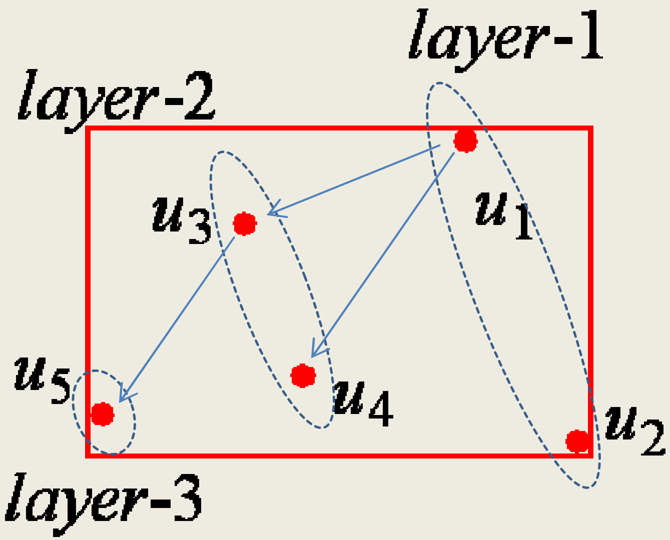
Layer Structure

- *layer*-1 is the skyline of all instances
- *layer*-*k* ($k > 1$) is the skyline of instances except those at *layer*-1, …, *layer*-(k-1)
- $\forall$ u at *layer*-*k*,
  $\exists$ u' at *layer*-(k-1),
  s.t., $u' \succ u$ and $Pr(u') \geq Pr(u)$
- $\max\{Pr(u) \mid u$ is at *layer*-(k-1)$\}$
  $\geq \max\{Pr(u) \mid u$ is at *layer*-$k\}$



Bounding with Layer Structures

- $\max\{Pr(u_1), Pr(u_2)\}$
  $\geq \max\{Pr(u_3), Pr(u_4)\}$
  $\geq Pr(u_5)$



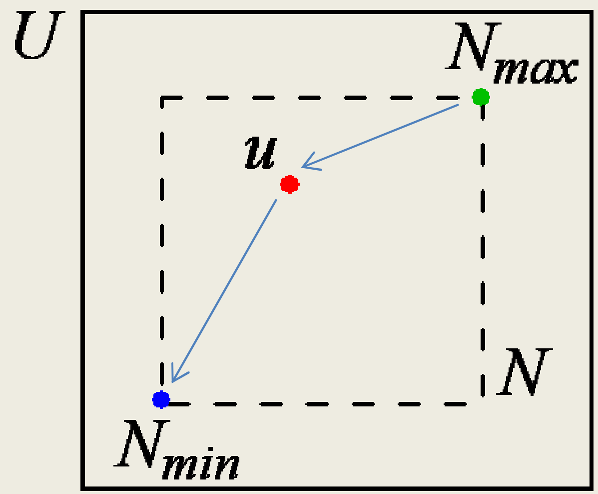
- Order: $u_1 \rightarrow u_2 \rightarrow u_3 \rightarrow u_4 \rightarrow u_5$

	u_1	u_2	u_3	u_4	u_5	U
lower bound	$Pr(u_1)$	$Pr(u_2)$	0	0	0	$(Pr(u_1)+Pr(u_2))/5$
upper bound	$Pr(u_1)$	$Pr(u_2)$	$Pr(u_1)$	$Pr(u_1)$	$Pr(u_1)$	$(4Pr(u_1)+Pr(u_2))/5$

- Compute $Pr(u_i)$
  - Build an R-tree for the instances of each object, traverse R-trees

Top-Down Method

- Lemma
  - Two instances u_1 and $u_2 \in U$, if $u_1 \succ u_2$, then $Pr(u_1) \geq Pr(u_2)$
  - N is a subset of instances of U, $\forall u \in N, Pr(N_{max}) \geq Pr(u) \geq Pr(N_{min})$



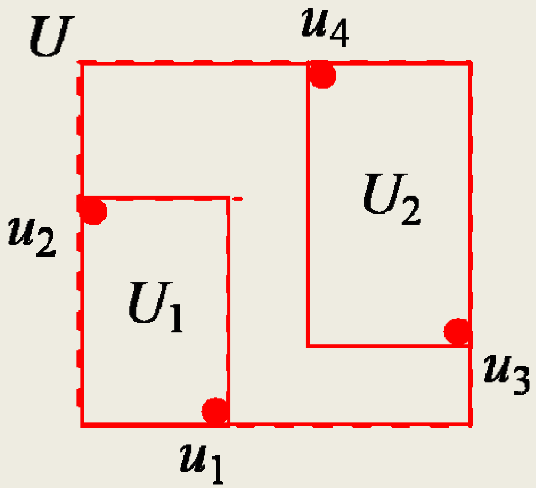
- Object U has l partitions $N1, \ldots, N_l$,

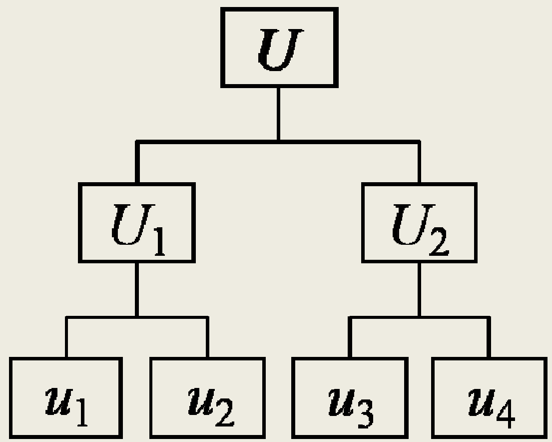
$$\frac{1}{|U|}\sum_{i=1}^{k}|N_i|\cdot Pr(N_{i,max}) \geq Pr(U) \geq \frac{1}{|U|}\sum_{i=1}^{k}|N_i|\cdot Pr(N_{i,min})$$

- Build a partition tree for each object to organize partitions

Partition Tree

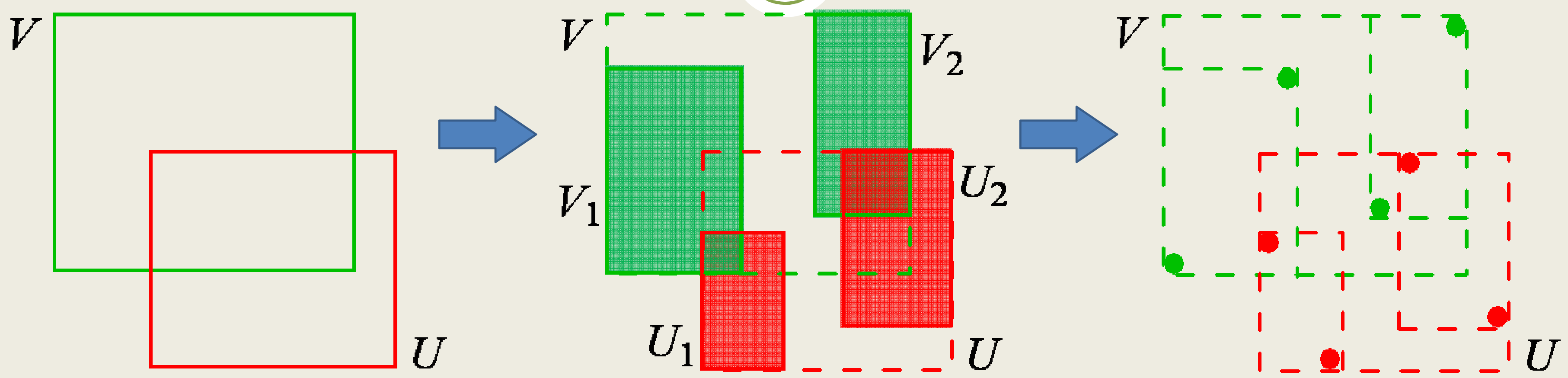
- Binary tree





- Growing one level of the tree in each iteration
  - Choose one dimension in a round-robin fashion
  - Each leaf node is partitioned into two children nodes, each of which has half of instances
- Bound $Pr(N_{max})$ and $Pr(N_{min})$ of a partition N

Bounding with Partition Trees



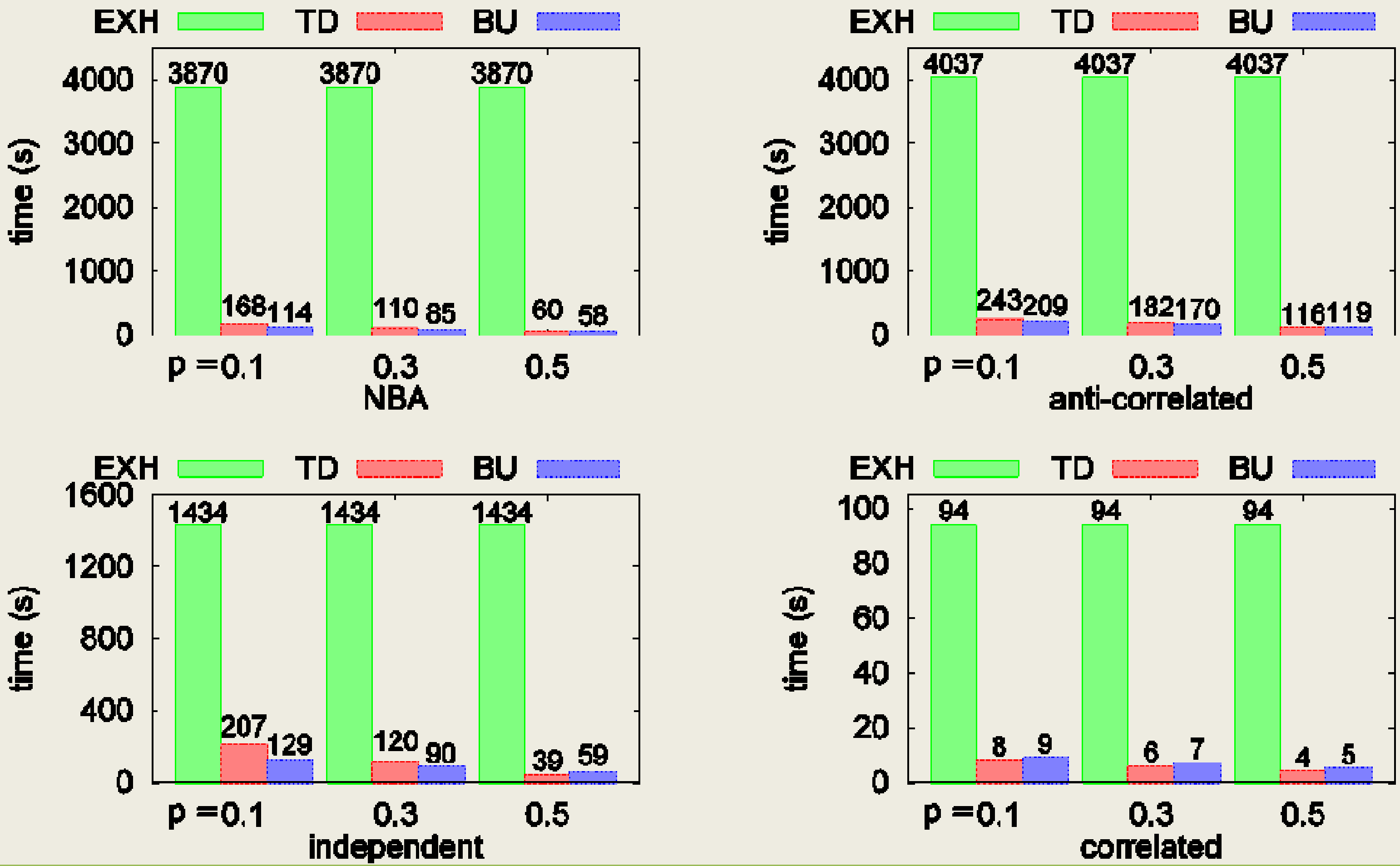
- Case 1: $V_{2,min} \succ U_{1,max}$, all instances in V_2 dominate $U_{1,max}$ and $U_{1,min}$
- Case 2: $V_{1,max} \nsucc U_{2,min}$, no instance in V_1 dominates $U_{2,max}$ or $U_{2,min}$
- Case 3: V_1 and U_1 are not in Case 1 or 2, do estimation
  - No instance in V_1 dominate $U_{1,max}$ – upper bound of $Pr(U_{1,max})$
  - All instances in V_1 dominate $U_{1,min}$ – lower bound of $Pr(U_{1,min})$
- $\frac{1}{|U|}\sum_{i=1}^{2}|U_i|\cdot Pr(U_{i,max}) \geq Pr(U) \geq \frac{1}{|U|}\sum_{i=1}^{2}|U_i|\cdot Pr(U_{i,min})$

Experiment Settings

- NBA data set
  - 339,721 game records of 1,313 players from 1991 to 2005
    259 records per player on average
  - 3 attributes: number of points, assists, rebounds
- Synthetic data sets
  - Distributions: anti-correlated, independent, correlated
  - Dimensionality: 2 ~ 10
  - Cardinality: 2,000 ~ 20,000
  - Average number of instances per object: 200
- Algorithms
  - Bottom-up algorithm and Top-down algorithm
  - Exhausted algorithm

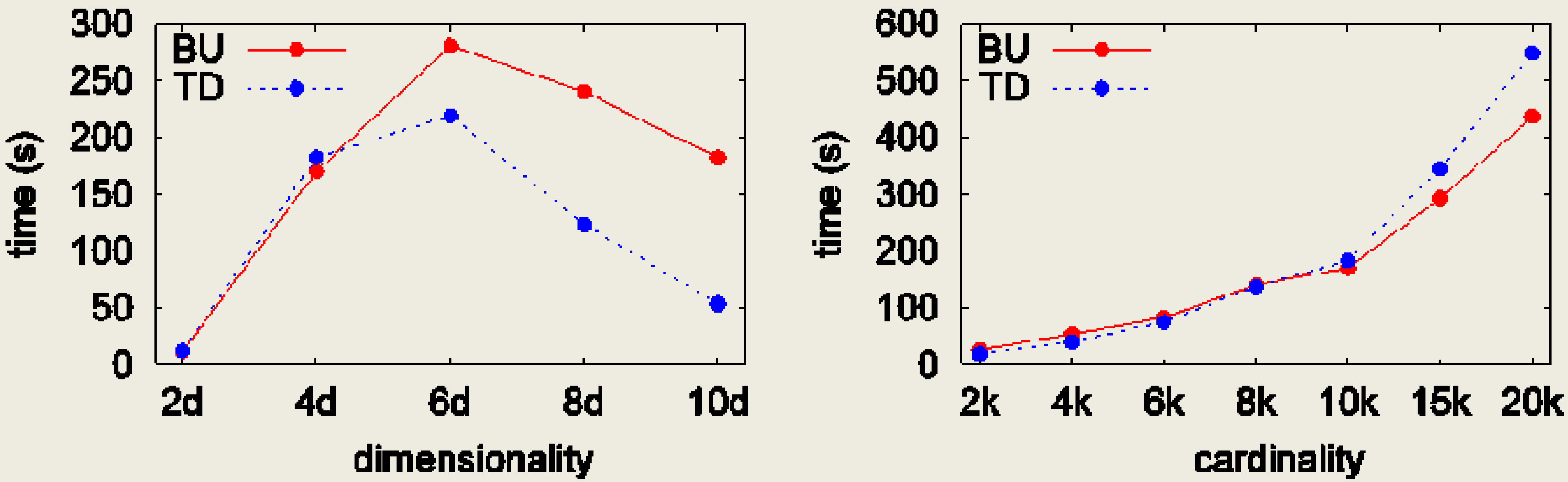
Overall Performance

- 4d-space 10,000 objects with average 200 instances each



Comparison of Two Algorithms

- Threshold $p = 0.3$



Conclusions

- Probabilistic skyline model
  - An object takes a probability to be in the skyline
- Two algorithms
  - Bottom-up
  - Top-down
- Experiments
- Future Work
  - Continuous case

Thank You!