On Skyline Groups

Chengkai Li[1], Nan Zhang[2], Naeemul Hassan[1], Sundaresan Rajasekaran[2], Gautam Das[1,3]
[1]University of Texas at Arlington, [2]George Washington University, [3]Qatar Computing Research Institute

MOTIVATION

- Find a group of experts
  - Form a team in online fantasy games
  - Perform a task in developing software
  - Group of reviewer for paper review
- Find a group of objects
  - Buying shares from Stock Market.

SKYLINE GROUP

	Points	Assists	Blocks
P_1	3	4	5
P_2	4	2	3
P_3	4	5	3
P_4	2	1	2
P_5	4	1	2

NBA Players Score

$D = \{P_1, P_2, P_3, P_4, P_5\}$

Find skyline groups of 3 players

	SUM			MIN			MAX		
	Points	Assists	Blocks	Points	Assists	Blocks	Points	Assists	Blocks
P_1, P_2, P_3	11	11	11	3	2	3	4	5	5
P_1, P_2, P_4	9	7	10	2	1	2	4	4	5
P_1, P_2, P_5	11	7	10	3	1	2	4	4	5
P_1, P_3, P_4	9	10	10	2	1	2	4	5	5
P_1, P_3, P_5	11	10	10	3	1	2	4	5	5
P_1, P_4, P_5	9	6	9	2	1	2	4	4	5
P_2, P_3, P_4	10	8	8	2	1	2	4	5	3
P_2, P_3, P_5	12	8	8	4	1	2	4	5	3
P_2, P_4, P_5	10	4	7	2	1	2	4	2	3
P_3, P_4, P_5	10	7	7	2	1	2	4	5	3

PROBLEM

Given a database table *D* of n tuples $\{t_1, t_2, \ldots, t_n\}$ and *m* attributes $A_1, A_2, \ldots, A_m$, a subset of *k* tuples i.e. $G : \{t_{i1}, \ldots, t_{ik}\}$ is a *k-tuple* group.

The problem is to find the skyline of *k-tuple* groups of *D* i.e. *Sky(D, k)*.

The groups are compared by their aggregates.

We consider Summation (SUM), Minimum (MIN) and Maximum (MAX) as aggregate functions.

SYSTEM FRAMEWORK

Baseline

n tuples, Group size *k* → Group Enumeration SUM/ MIN/ MAX → $\binom{n}{k}$ combinations → Skyline Operation → All Skyline Groups

Challenges

- $\binom{n}{k}$ is very large number, we may not be able to afford computing or storing such big number of group combinations.
- Number of all skyline groups can be very large.

Our Framework

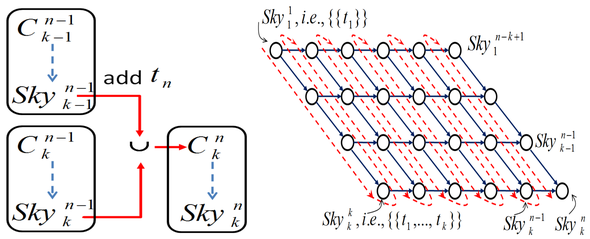
n tuples, Group size *k* → Input Pruning → n' where $n >> n'$ → Search Space Pruning (OSM/WCM) → Candidate Groups, Candidate Groups << $\binom{n'}{k}$ << $\binom{n}{k}$ → Skyline Operation Output Pruning → Unique Skyline Vectors → Post Processing → All Skyline Groups

SEARCH SPACE PRUNING: OSM

P_2, P_3, P_5 → a 3-tuple skyline of D.
P_2, P_3 → a 2-tuple skyline of D – $\{P_5\}$.
P_2 → a 1-tuple skyline of D – $\{P_3, P_5\}$.

We arbitrarily order tuples as $D = \{t_1, t_2, \ldots, t_n\}$ and call *k-tuple* skyline as Sky_k^n. From above observation, if $G \in Sky_k^n$ and $t_n \in G$, then $G \backslash \{t_n\} \in Sky_{k-1}^{n-1}$.

We develop a dynamic programming based algorithm based on this property and name it Order Specific Property Method (OSM).



This property is satisfied by SUM and also can be extended for MIN and MAX.

OUTPUT PRUNING

- Multiple groups with the same score.
- Observed in MAX and MIN but rare in SUM.
- Find unique skyline vectors instead of all skyline groups.

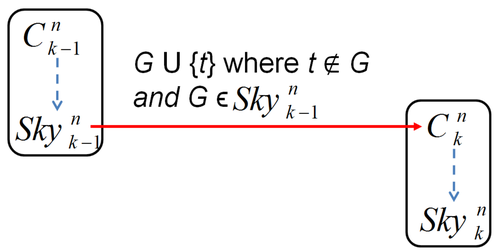
For instance,

$\{P_1, P_2, P_3\}$, $\{P_1, P_3, P_4\}$ and $\{P_1, P_3, P_5\}$ are all skyline groups for MAX function having same unique skyline vector [4, 5, 5].

SEARCH SPACE PRUNING: WCM

For MIN and MAX, if $G \in Sky(D,k)$, then at least one *(k-1)-tuple* subset of *G* will be a *(k-1)-tuple* skyline. This is weaker than Apriori property; such called Weak Candidate Generation Property Method (WCM).

This is only valid when distinct value property holds. However, our algorithm extends this to general cases.

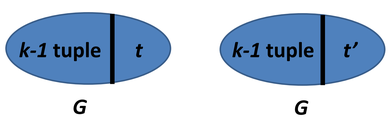


WCM is satisfied by MIN and MAX. SUM does not satisfy this property.

INPUT PRUNING

If a tuple is dominated by >= k tuples, it can be discarded for unique skyline vector calculation.

	Points	Assists	Blocks
P_1	3	4	5
P_2	4	2	3
P_3	4	5	3
P_4	2	1	2
P_5	4	1	2



G' dominates G

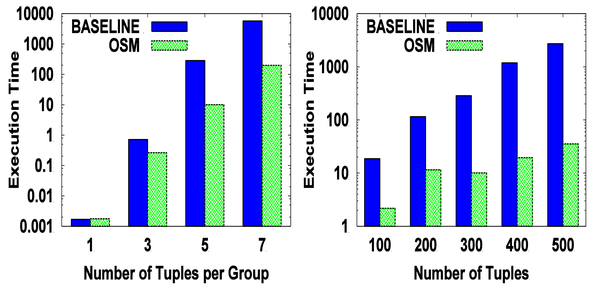
Input pruning significantly reduces size of *n*.

POST PROCESSING

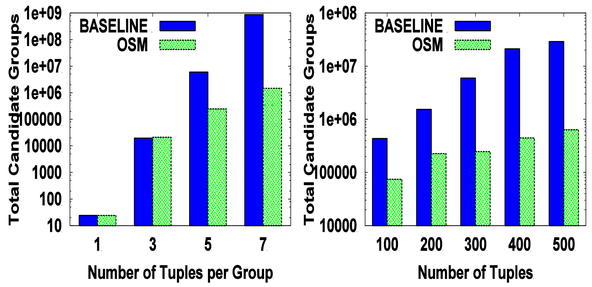
This step is necessary for finding all skyline groups from unique skyline vectors

- **MIN**: It is sufficient to find all input tuples which are equal to or dominate a skyline vector and then find *k-tuple* combination of these; time complexity O(n).
- **MAX**: The problem is NP-hard. But simple brute-force is practically efficient because of small input size.

EXPERIMENTS



Execution time comparison of Baseline and OSM for SUM function. Time in seconds, logarithmic scale.



Number of generated candidate groups comparison of Baseline and OSM for SUM function. Logarithmic scale.

Data Set

- NBA players performance statistics (points, assists, rebounds, steals and blocks per game) in 2009 regular season.
- Synthetic Dataset of 10 Million records of 5 attributes.

REFERENCES

[1] C. Li, N. Zhang, N. Hassan, S. Rajasekaran, and G. Das. On Skyline Groups. In *CIKM* 2012.