On "One of the Few" Objects

You Wu, Pankaj K. Agarwal, Jun Yang
Dept. of Computer Science
Duke University, USA
{wuyou, pankaj, junyang}@cs.duke.edu

Chengkai Li
Dept. of Computer Science and Engineering
The University of Texas at Arlington, USA
cli@uta.edu

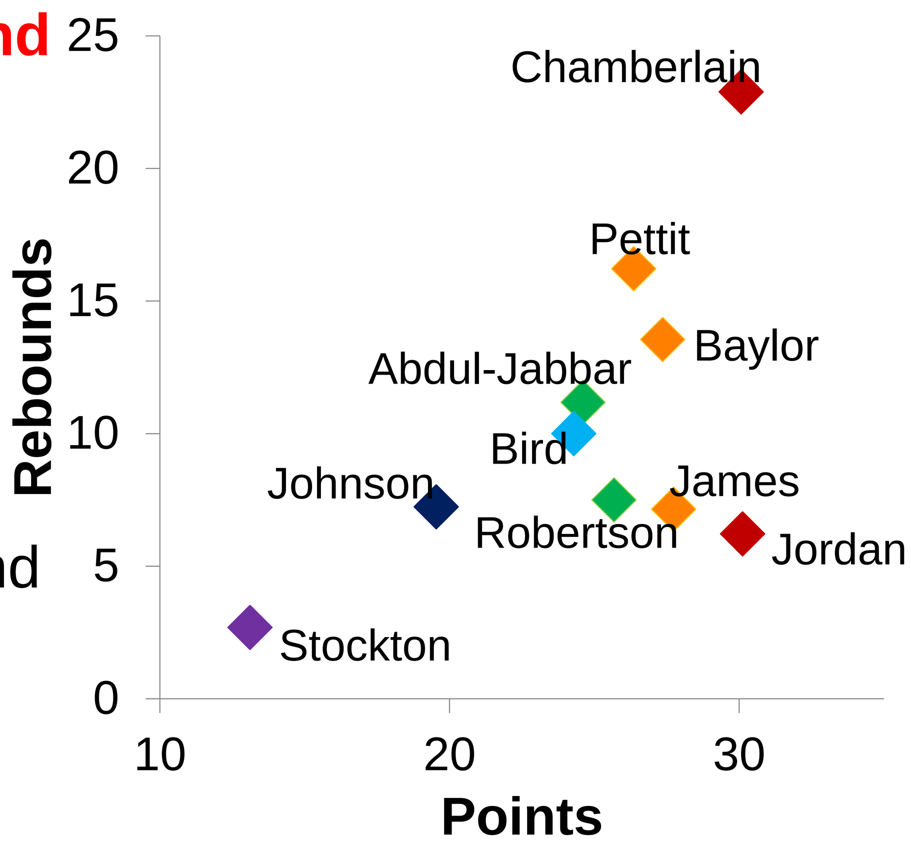
Cong Yu
Google Research
congyu@google.com

"One of the Few" Statements

- "Karl Malone is **one of the only two** players in NBA history with at least 25,000 points, 12,000 rebounds, and 5,000 assists in one's career"
- …implies "**only one** other player (Kareem Abdul-Jabbar) **dominates** Karl Malone in career total points, rebounds, and assists"
- Applications in **computational journalism**: e.g., fact checking and finding

Connection with Skyline/Skyband

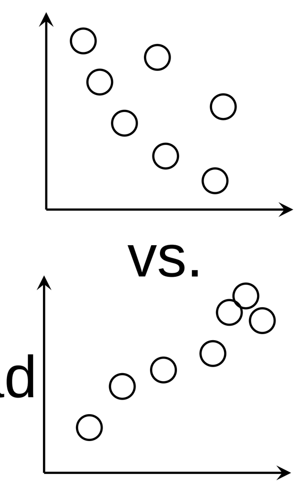
- Object in "**one of the k**" statement ⇔ **dominated by less than k** other objects ⇔ point in the **k-skyband** (all referring to the same "subspace," e.g., {points, rebounds})
- E.g. **Bird** in 4-skyband ⇔ he can be stated "one of the 5"



Objectives, Challenges, and Solutions

Ensure "interestingness" of a statement

- Small k in "one of the k" statement ≠ interesting!
  - Many objects can be "one of the k," especially in high dimensions
- Letting user handpick k is tricky: it depends on not only dimensionality but also data distribution
- **Limit the size τ of the skyband** instead → **top-τ skyband**
  - Simple to interpret: the same statement can be made for $\leq \tau$ objects in the same subspace
  - Applicable to all subspaces: # of tiers (k) in the skyband automatically adapts to size (τ)



Find all interesting statements for all subspaces

- Efficiency within each subspace: **avoid computing a huge skyband with size $> \tau$**
- Efficiency across subspaces: **share computation; prune subspaces**

Rank objects by their "interestingness"

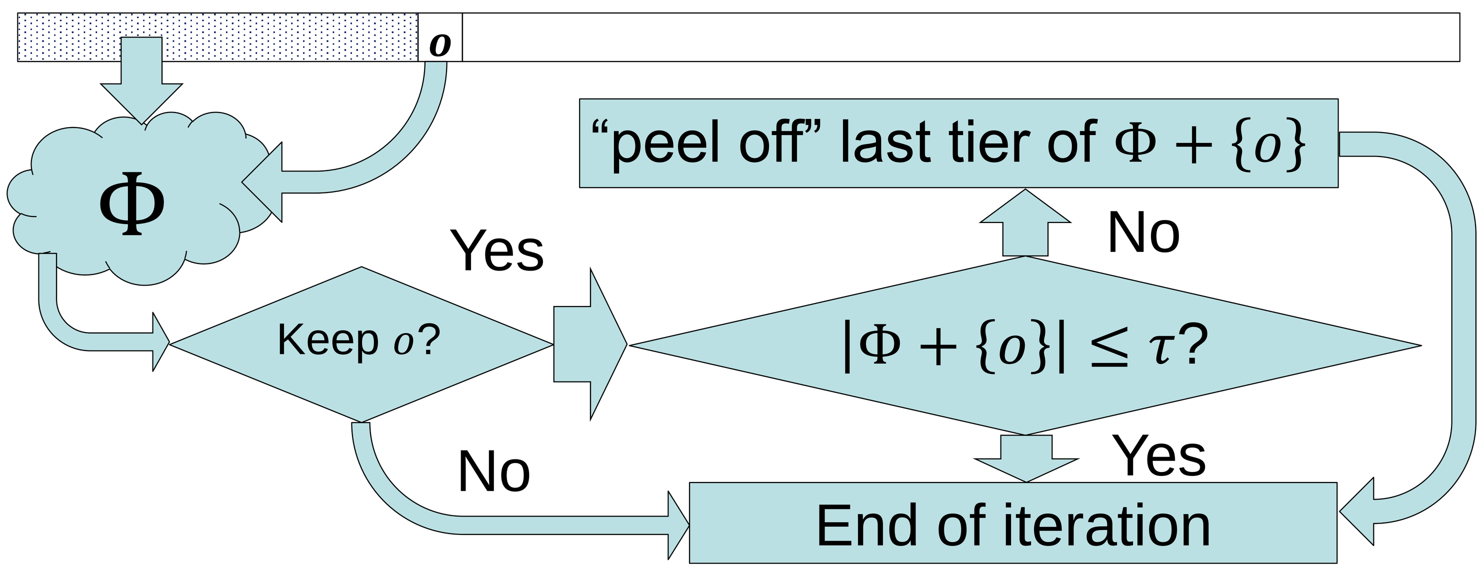
x Score and rank objects using a weight vector
  - Tough to tune: # knobs = dimensionality

x Kemeny-optimal rank aggregation over all single-dimension rankings
  - Expensive to compute
  - Cannot customize: who is better—a specialized player like John Stockton (2nd in "assists", 404th/1622nd in "points"/"rebounds") or a well-rounded one like Charles Barkley (33rd/23rd/223rd in "points"/ "rebounds"/"assists")

✓ **A ranking method that is easy to use & compute**
  - Aggregates "interestingness" across all subspaces
  - Provides a single knob (α) to tune preference towards specialized vs. well-rounded objects

Computing Top-τ Skyband in a Subspace

- Visit objects in a "safe" order, e.g. topological order
- $O(\tau)$ time each iteration



- Efficient for large n, smaller τ
  - <10mins on NBA dataset with $n \approx 400k, \tau = 100$
  - Takes brute force approach days to complete

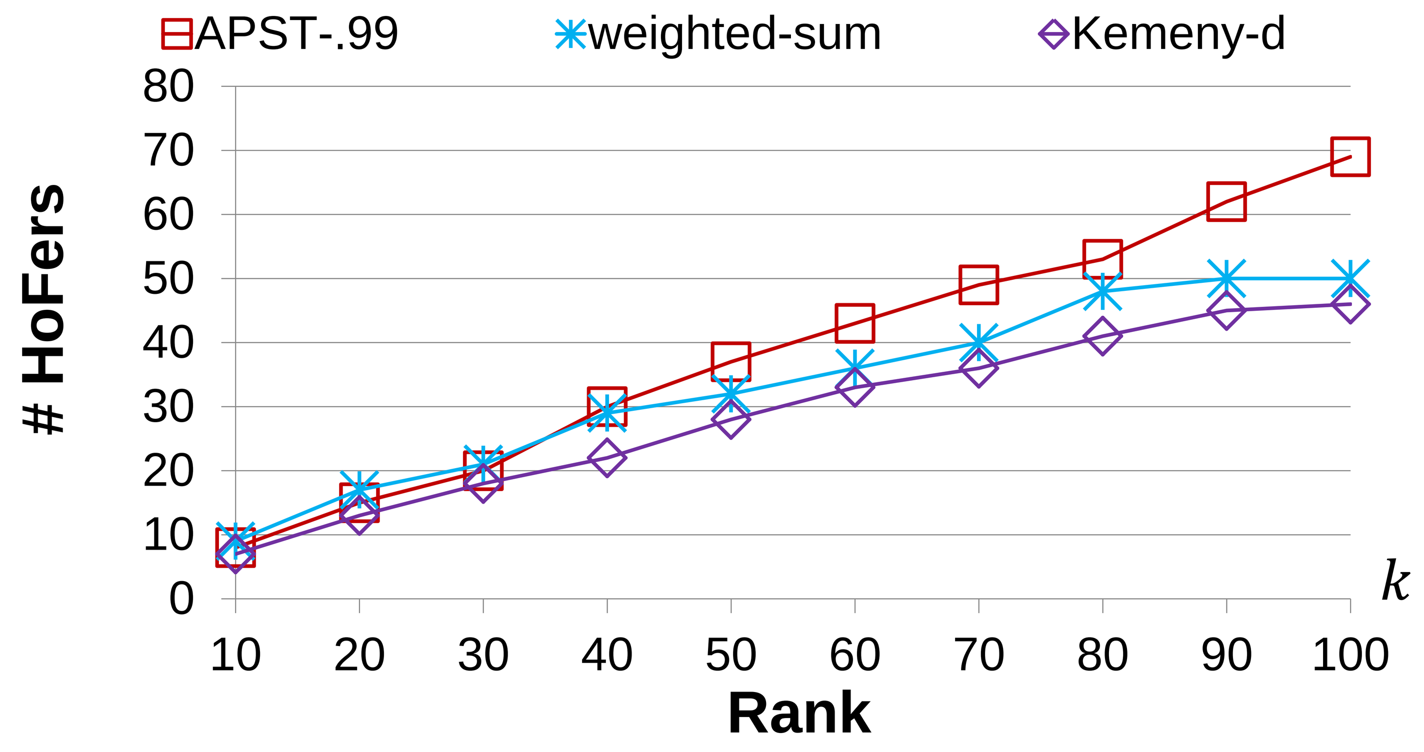
Ranking Method (APST-α)

- In each subspace, assign a decaying score (by α^{i-1}) according to rank (i)
  - Ties (those in the same tier of the skyband) share their total score
- Across all subspaces, add scores for the same object
- E.g. scores for Baylor with
  - 4th in {Points} ⇒ 2^{-3}
  - 3rd in {Rebounds}⇒ 2^{-4}
  - Tied with James, Pettit at 3rd, 4th, 5th in {Points, Rebounds} ⇒ $(2^{-2} + 2^{-3} + 2^{-4})/3$
  - Total for ranking = $2^{-3} + 2^{-4} + (2^{-2} + 2^{-3} + 2^{-4})/3$

Rank	Score		Object
1	2^{0}		Chamberlain
2	2^{-1}		Jordan
3	2^{-2}	avg	Baylor
4	2^{-3}		James
5	2^{-4}		Pettit
6	2^{-5}	avg	Abdul-Jabbar
7	2^{-6}		Robertson
8	2^{-7}		Bird
9	2^{-8}		Johnson
10	2^{-9}		Stockton

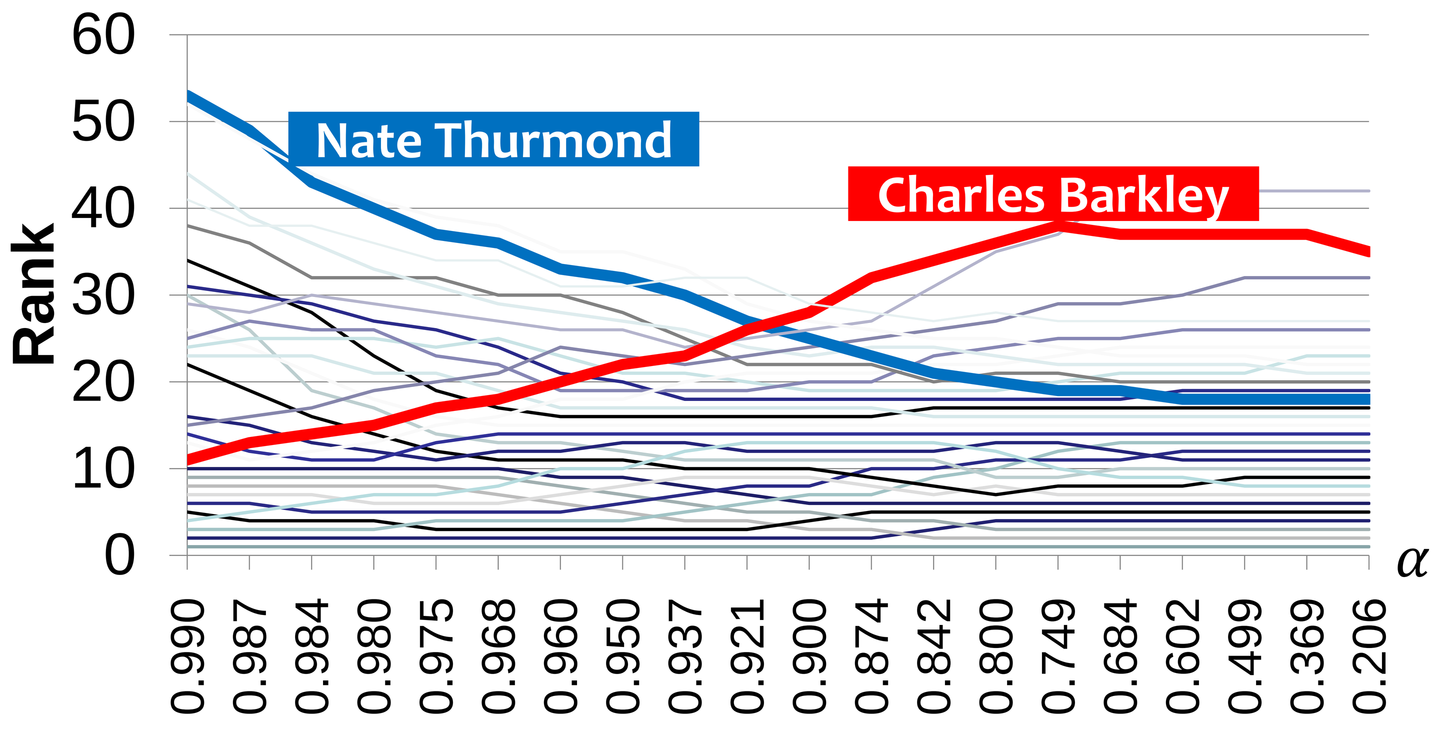
Effectiveness of Ranking

Identifying Hall of Fame NBA players (HoFers)



- # HoFers identified by 3 ranking methods in top-k
- Higher precision by APST-α
  - Similar performance for different α

Tuning specialized vs. well-rounded objs with α



- As α decreases,
  - rank of **well-rounded** object drops
  - rank of **specialized** object rises

Paper ID: 843